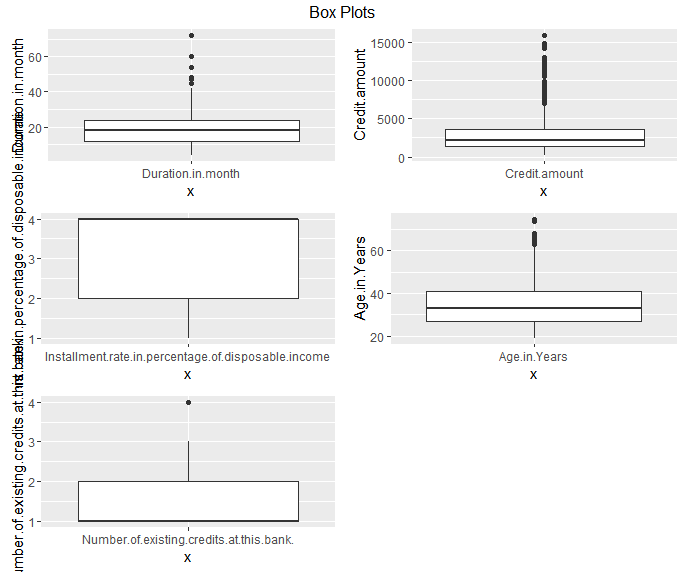
LOGISTIC REGRESSION SUBMISSION

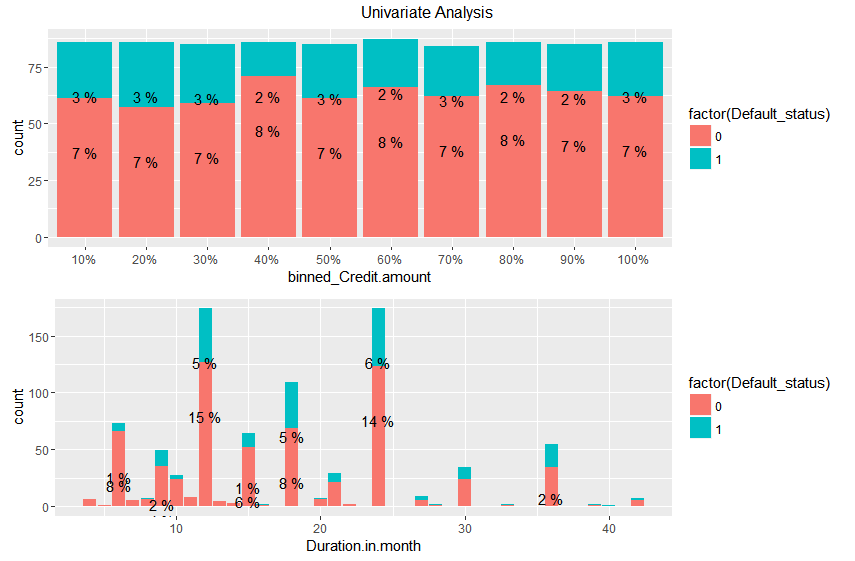
**NOTE:** This should briefly describe the important results and recommendations. The structure is suggestive; make sure to not exceed 7 pages**.**

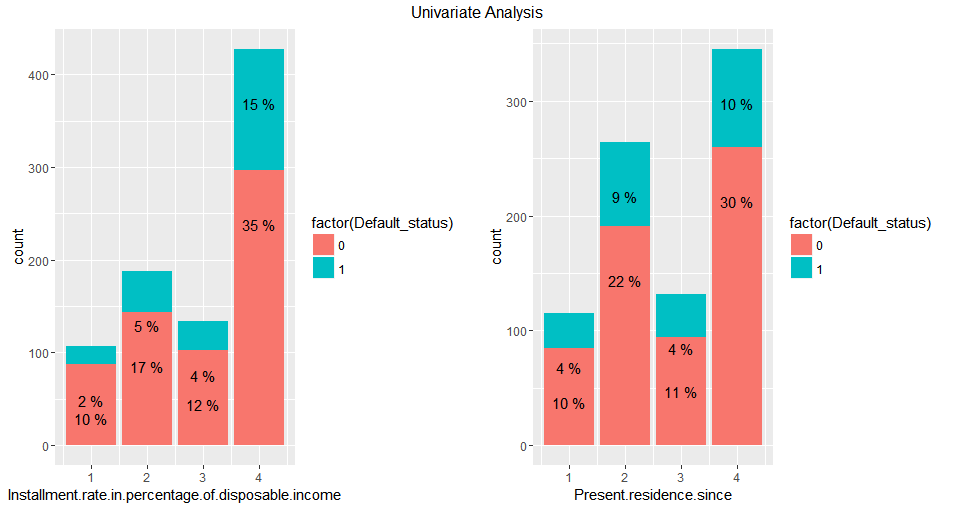
# Checkpoint-1: Data Understanding and Data Exploration

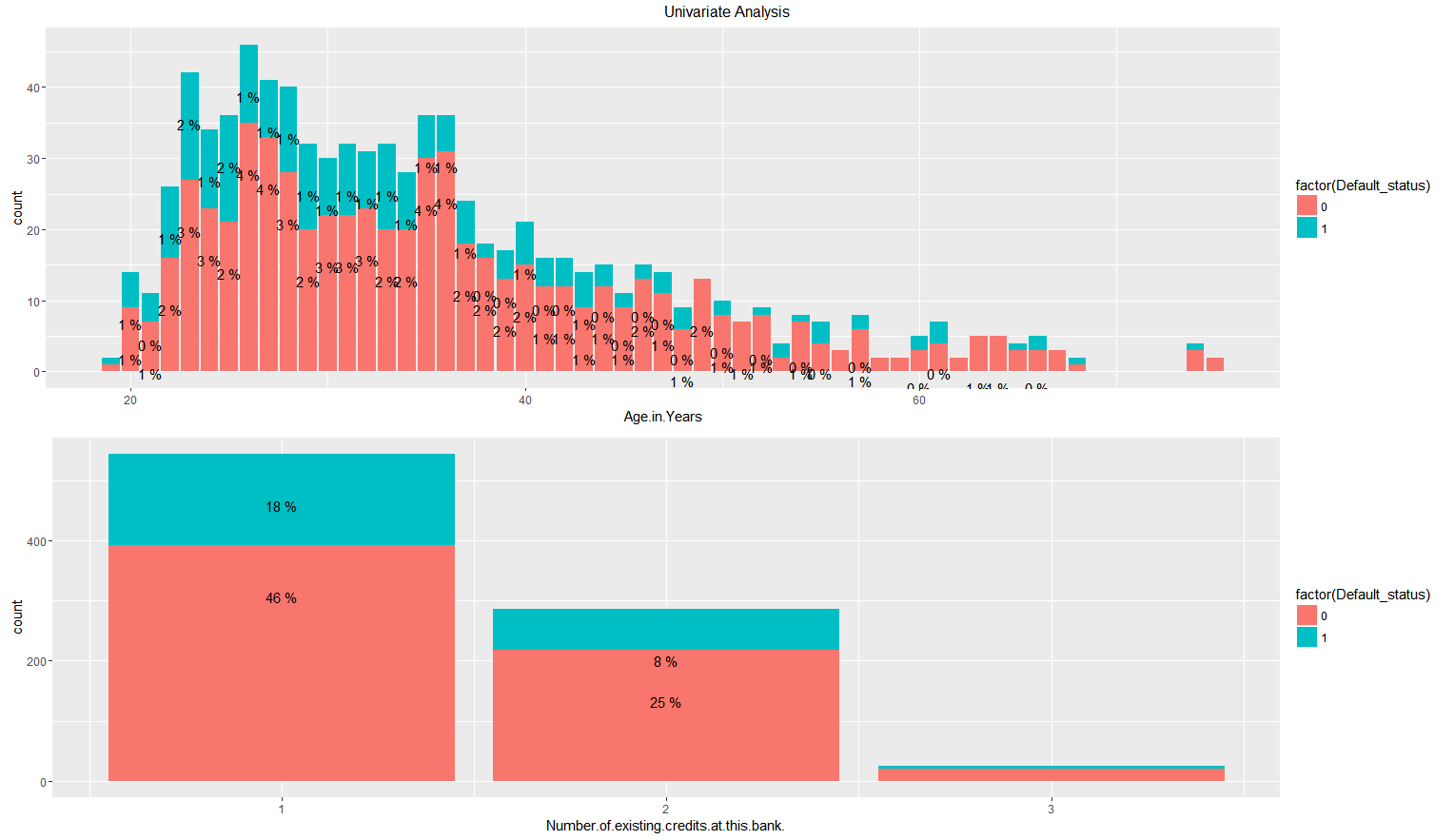
* Display the plots and explain the insights



Outliers can cause problems with Model , so we explore data for Outliers with box plots also gives us good pic of means . So there are Outliers which need to be taken care of.



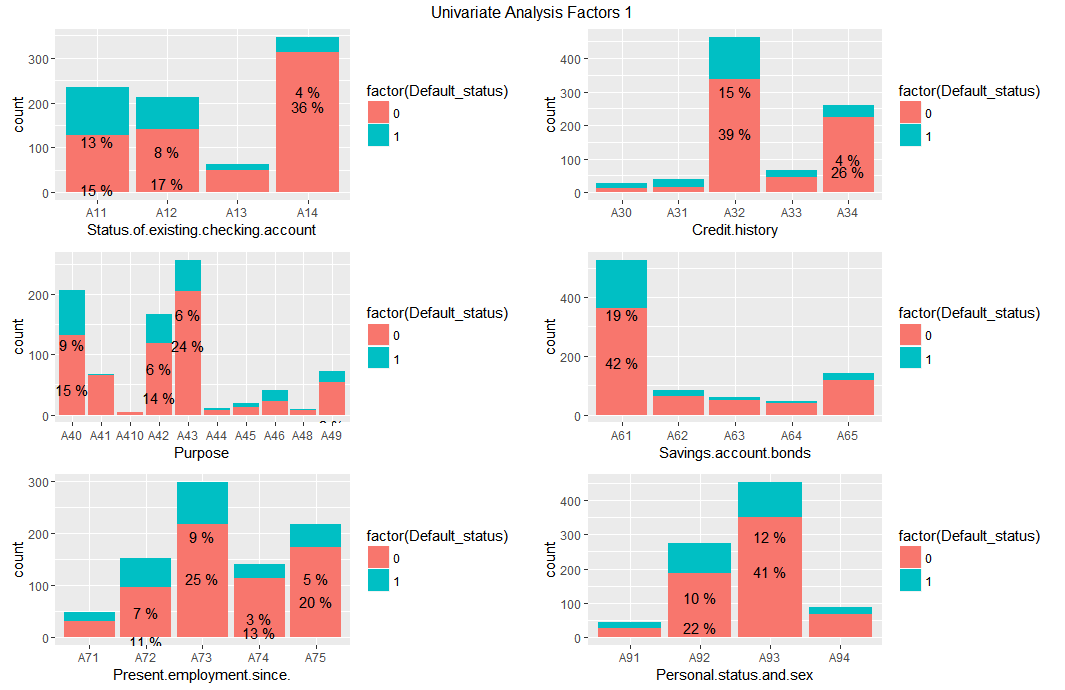




Maximum Defaults (Continues variables):

Number Existing credits : 1  
Age : 20-40

Instalment in % of income : 4





# High Defaulters: A11 : Status of existing checking account < 0 DM A40+A42 : Loan for Car , furniture/equipment A61 : Savings account/bonds < 100 DM A73 : 1 to 4 years Present employment since A101: No Other debtors / guarantors A143: Other instalment plans : None A152: Own House Number of Existing credits in Bank is = 1 A191: Telephone None

# 

# 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:   
  Here outlier was removed As we had sufficient data also imputing outliers means effecting co-related variables. 856 Obs ,
* Explain the methodology of how did you created dummy variables :   
  Dummy variables were created for factor variables , using model.matrix command and then merging them back to data frame. 47 Variables
* If binning for numerical variables done explain why it was required?   
  Probably Credit amount could have been binned, but I didn't play with that

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | Duration.in.month,  Credit.amount, Number.of.existing.credits.at.this.bank. |
| Dummy creation | Status.of.existing.checking.account,  Credit.history,  Purpose,  Savings.account.bonds, Present.employment.since.,  Personal.status.and.sex, Other.debtors...guarantors,  Property,  Other.installment.plans,  Housing.,  Number.of.existing.credits.at.this.bank.,  Telephone., foreign.worker |
| Binning of variables | None |

# Checkpoint 3: Splitting the Dataset into train and test

Training set 70% , german\_credit\_train  
 Testing set 30% , german\_credit\_test

Using R function sample.split with seed = 100

# 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  
  We Start with making sure all variables are numeric by creating dummy for factors etc   
  then we run step function which will give the least AIC , Now try and remove variables which are non significant.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeeric)** |
| Duration.in.month | 1.22e-05 |
| Age.in.Years | 0.105485 |
| Status.of.existing.checking.accountA12 0to200DM | 0.006248 |
| Status.of.existing.checking.accountA13 >200DM | 0.080133 |
| Status.of.existing.checking.accountA14 (no checking account) | 4.95e-08 |
| Credit.historyA32 existing credits paid back duly till now | 0.012453 |
| Credit.historyA34 critical account | 9.45e-06 |
| PurposeA41 car (used) | 0.000486 |
| PurposeA42 furniture/equipment | 0.003256 |
| PurposeA43 radio/television | 0.000169 |
| PurposeA49 business | 0.016094 |
| Savings.account.bondsA65 No savings account | 0.023202 |
| Present.employment.since.A74 4 to 7 | 0.048358 |
| Personal.status.and.sexA93 (male : single) | 0.076448 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 583.02 |
| Null deviance | 691.21 |
| Residual Deviance | 553.02 |

# Checkpoint 5: Model Evaluation

* Calculate c-statistic and KS-statistic. What can you tell about the model based on their values?  
    
   This model has 89 % accuracy on training set and 87% accuracy on test set. Hence, it can be assumed that the model will do fair job of predicting Credit Default.

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 | 0.7968799 | C-statistic | 0.8035325 |
| KS-statistic | 0.47 | KS-statistic | 0.498 |
| Model Evaluation (write Accept or Reject) | | @ 80% accuracy with > 0.4 descretion we can accept the model | |

# Checkpoint 6: Threshold value

* Select an appropriate threshold value and calculate the confusion matrix and overall accuracy, sensitivity and specificity

Selected Threshold is probability = **0.35 [see graph below] shortest distance from (x=0,y=1)**

**min( sqrt( (x-0)^2 + (y-1)^2 ) )**

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.7626 |
| Sensitivity | 0.72 |
| Specificity | 0.77 |

