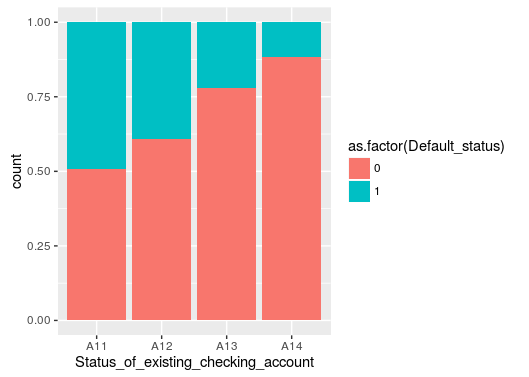
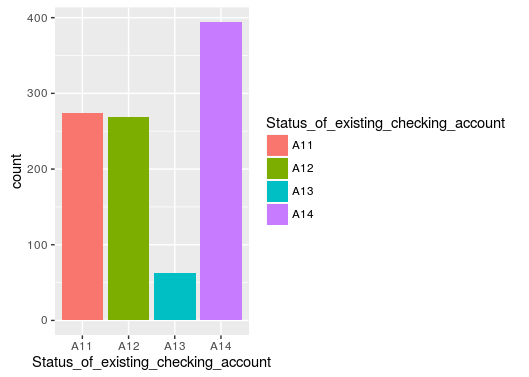
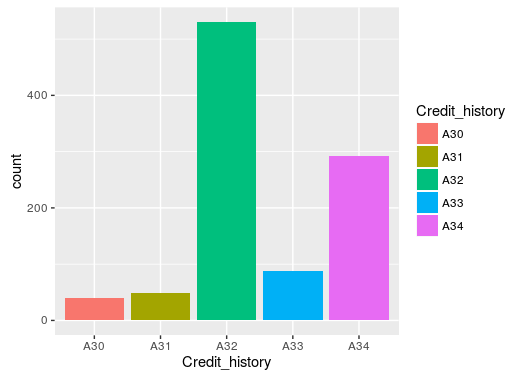
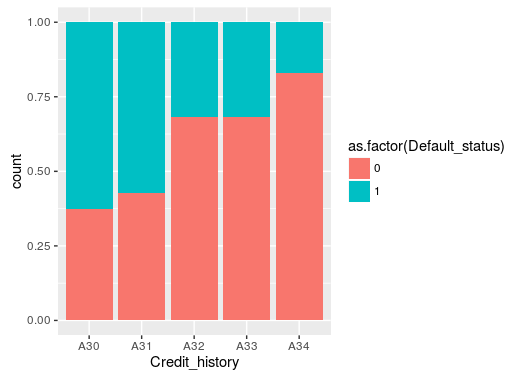
LOGISTIC REGRESSION SUBMISSION

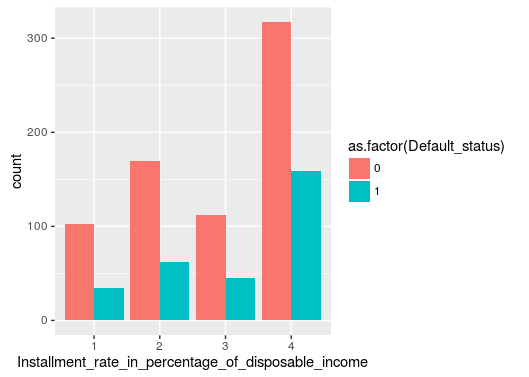
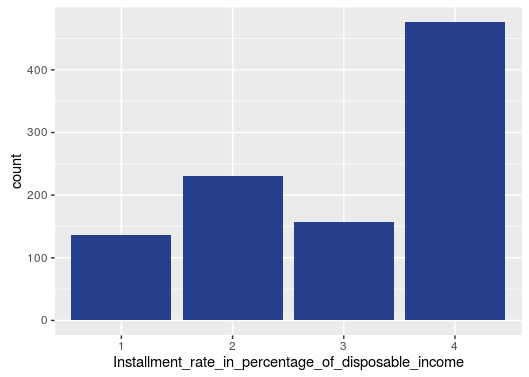
# Checkpoint-1: Data Understanding and Data Exploration



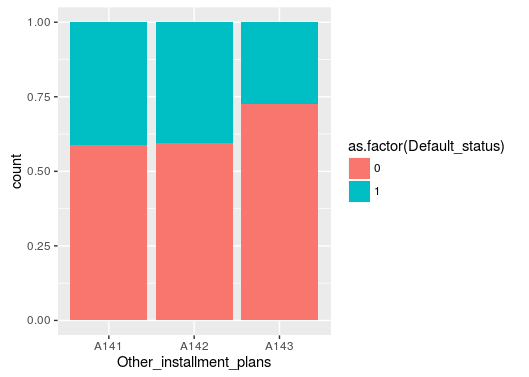
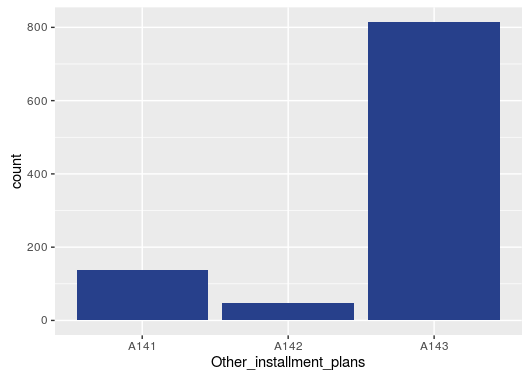
* Most loans have no checking account
* The Default rate is least for accounts with no Checking Account(
* The Default Rate is higher for account with >200DM(A13).

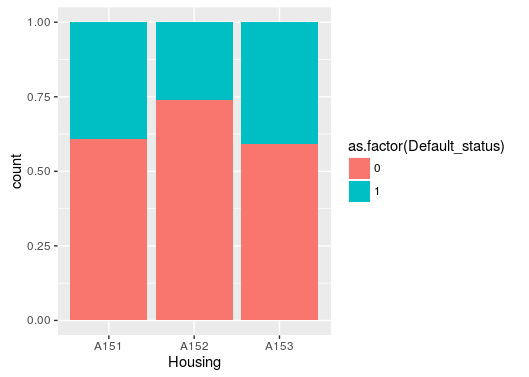
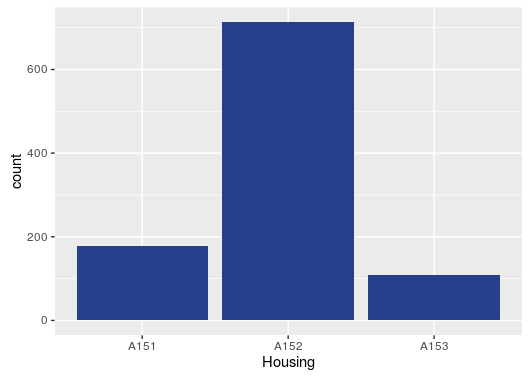
* Most loans have credit history as “Existing Credits, paid back duly” (A32)
* Default Rate is least for customers, who have credits in other banks
* A33 and A34 have almost same default rates.



* Most Credits, have instalment Rate of 4%.
* The highest default Rates are for credits with instalment Rate of 4%



* Most Credits, are for customers who have no other instalment plans (A143)
* The Default Rate is least for customers with no other instalment plans.



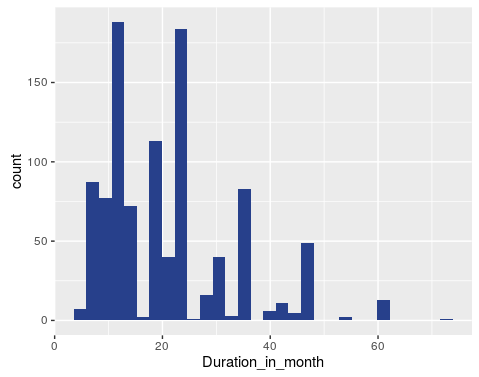
* Most customers, who take credit have own house (A152) and have least default rate

# 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 | 22 |
| Total missing values in the dataset | No Missing Variables |

* Outlier Treatment
  + Credit\_Amount : log transformation was first done, and boxplot and quantile distribution was observed. We found that there was a jump is log(Credit\_Amount) from 0% to 1%. So we floored the transformed variable to 6.054435
  + Age\_in\_years : Used quantile function to remove outliers. 97.7 percentile of the customers are below 64 years of age, hence we cap the Age at 64.
* Creating Dummy Variables
  + Convert all the character columns to factor.
  + For each factor column apply function model.matrix(~x-1,data)[,-1] using sapply.
  + Convert the model\_matrix into dataframe variable dummies.
  + Cbind the non-factor columns in german\_credit dataframe with the dummies dataframe created above.
* Binning
  + Binning was done for duration. As from the histogram below, we can observe that for every 12 months, there is a peak. Also, between 6 to 12 months also, we see a surge in loans.



* + Duration\_in\_month : For this variable, I have binned it into 7 categories and created a new column Duration.The levels in Duration are "LessThan1Yr", "OneYr", "LessThan1.5Yr","LessThan2Yr" , TwoYr", "LessThanOrEqual3Yr" and "GreaterThan3Yr"

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | Age\_in\_years and Credit\_amount |
| Dummy creation | [1] "Status\_of\_existing\_checking\_account"  [2] "Credit\_history"  [3] "Purpose"  [4] "Savings\_accountorbonds"  [5] "Present\_employment\_since"  [6] "Installment\_rate\_in\_percentage\_of\_disposable\_income"  [7] "Personal\_status\_and\_sex"  [8] "Other\_debtors\_or\_guarantors"  [9] "Present\_residence\_since"  [10] "Property"  [11] "Other\_installment\_plans"  [12] "Housing"  [13] "Number\_of\_existing\_credits\_at\_this\_bank"  [14] "Job\_status"  [15] "Number\_of\_people\_being\_liable\_to\_provide\_maintanance\_for"  [16] "Telephone"  [17] "foreign\_worker"  [18] "Duration" |
| Binning of variables | Duration\_in\_month to create new variable Duration |

# Checkpoint 3: Splitting the Dataset into train and test

The dataset has been split into 70:30 ratio. The train dataset has 700 observations and test dataset has 300 observations. After data preparation, we have 60 independent variables.

The dependent variable is Default\_status. This is 1 for Defaulters and 0 for non-Defaulters. In our modelling, we will consider Default\_status=1 as the positive class.

# Checkpoint 4: Modelling

* We use train dataset for building our logisitic regression model.
* For logistic regression we use the function glm. The command to build the initial model is as follows
  + glm(Default\_status~.,data=train,family = "binomial")
  + The AIC of this model was 734.73
* We then apply stepwise selection, using the step() function to select the best variables for our model.
* This results in model with 25 variables, with an AIC of 684.68
* We use vif() function to check for multicollinearity.For our assignment, we consider the VIF tthreshold as 3. CreditHistory.xA32 is the only variable with VIF>3, but it is highly significant and removing this variable from our model, results in a rapid increase in AIC.
* We start removing the insignificant variables one by one, by removing the variables with high p value.
* After removing a few insignificant variables, we find that AIC of CreditHistory.xA32 has reduced to below 3 and the AIC of the model is 703.41
* This is my final model. It has 11 variables.

|  |  |  |
| --- | --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeric)** | **Explaination of the coefficients** |
| Intercept | 1.4621 |  |
| Status\_of\_existing\_checking\_account.xA13 | -0.9890 | If money in Savings Account is >200DM, Less chance of defaulting ,hence negative coefficient |
| Status\_of\_existing\_checking\_account.xA14 | -1.9414 | If there is no checking account, then Default is less |
| Credit\_history.xA32 | -1.3045 | If previous credits have been paid duly, then probability of defaulting is low.Hence the negative coefficient |
| Credit\_history.xA33 | -1.2408 | There is a delay is paying of previous credits, then coefficient is higher than that of CreditHistory.xA32, but still since he has paid back before, he will pay back |
| Credit\_history.xA34 | -1.8203 | There are other credits at other banks,then lesser chance of defaulting. |
| Installment\_rate\_in\_percentage\_of\_disposable\_income.x4 | 0.6285 | 4 is the highest intrest rate, which means, customers may have a problem paying back and hence more default rate |
| Personal\_status\_and\_sex.xA93 | -0.4350 | If you are male and single, means you have lesser responsibilities which means, higher probability of not defaulting |
| Present\_residence\_since.x2 | 0.5699 | If staying in a place for 2 years, then I have higher tendency to default |
| Other\_installment\_plans.xA143 | -0.5418 | If No other instalment plans, then paying off debt is easier and hence low default rate. |
| Housing.xA152 | -0.4958 | Having own homes, indicates stablility and hence lower default |
| Duration.xGreaterThan3Yr | 0.9231 | If credit tenure is greater than 3 years than Default rate is high. |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 703.41 |
| Null deviance | 855.21 |
| Residual Deviance | 679.41 |

# Checkpoint 5: Model Evaluation

* c-statistic
  + The value for train set is 7.972838e-01
  + The value for test set is 7.058730e-01
  + We can say that model is highly discriminative as it have a c statistic of greater than 0.70
  + C statistic is equal to area under ROC Curve.
* KS Statistic
  + The Ks Statistic is in 2nd decile for both train and test, with value 0.1928571 and 0.1314286
  + The ks statistic indicates the degree of separation of the positive and negative distribution.
  + This means, KS statistic is top 3 deciles indicates a good model..

|  |  |  |  |
| --- | --- | --- | --- |
| **Train Dataset** | | **Test Dataset** | |
| C-statistic | 0.7972838 | C-statistic | 0.7058730 |
| KS-statistic | 0.1928571 | KS-statistic | 0.1314286 |
| Model Evaluation (write Accept or Reject) | | Accept the Model | |

# Checkpoint 6: Threshold value

* The Threshold value, depends on whether we want to maximise accuracy or maximise sensitivity.
  + **For Maximising Sensitivity**, we must select a lower threshold, so that no defaulter can be misclassified.
    - The Threshold for Maximising Sensitivity is chosen as **0.16**
    - The Accuracy, Specificity and Sensitivity on the test data for this threshold is as below :

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.6133 |
| Sensitivity | 0.8444 |
| Specificity | 0.5143 |

* + **For Maximising Accuracy**, the threshold selected is **0.64**
    - The Accuracy, Specificity and Sensitivity for this threshold on test data is as below:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.72 |
| Sensitivity | 0.17778 |
| Specificity | 0.95238 |