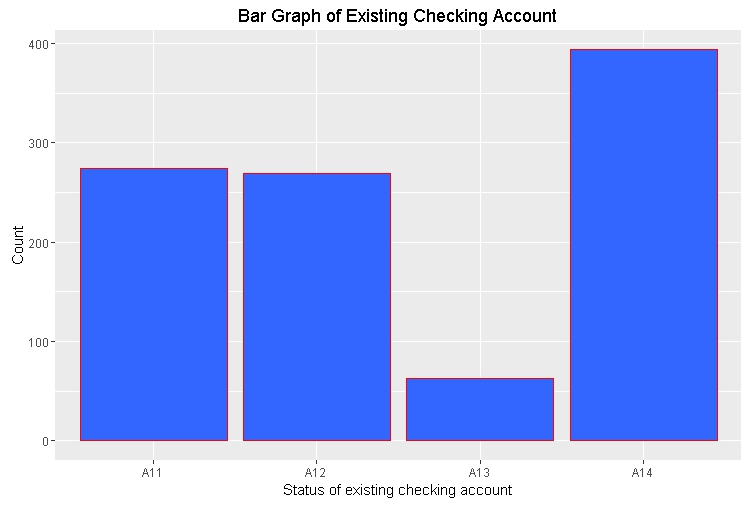
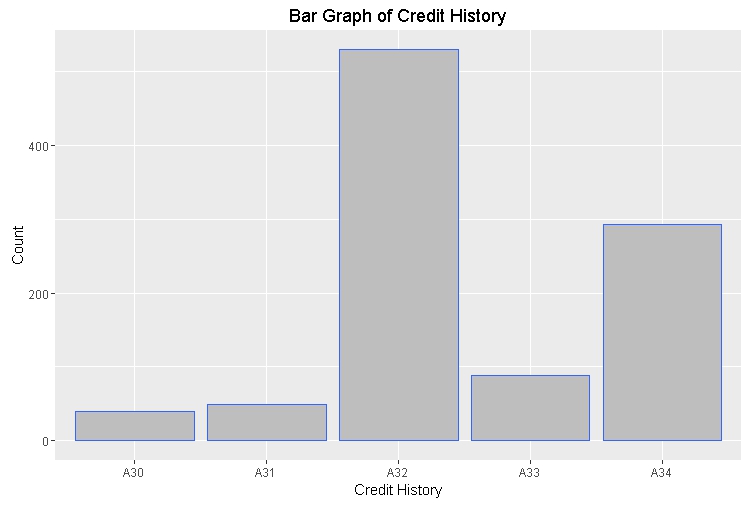
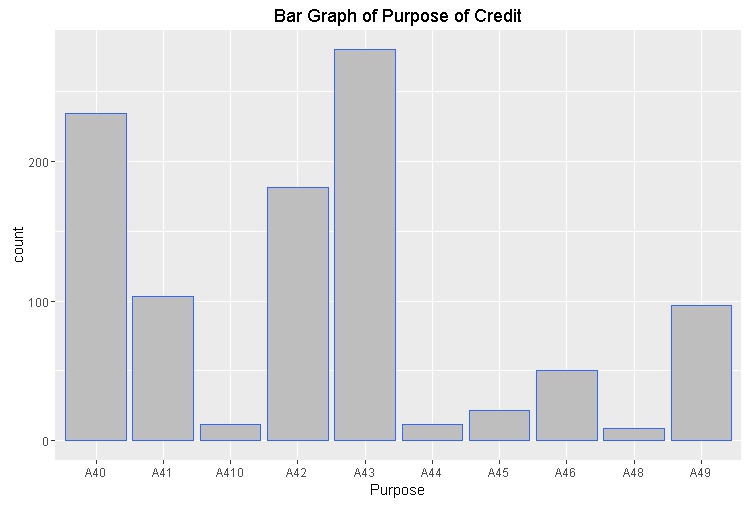
LOGISTIC REGRESSION SUBMISSION

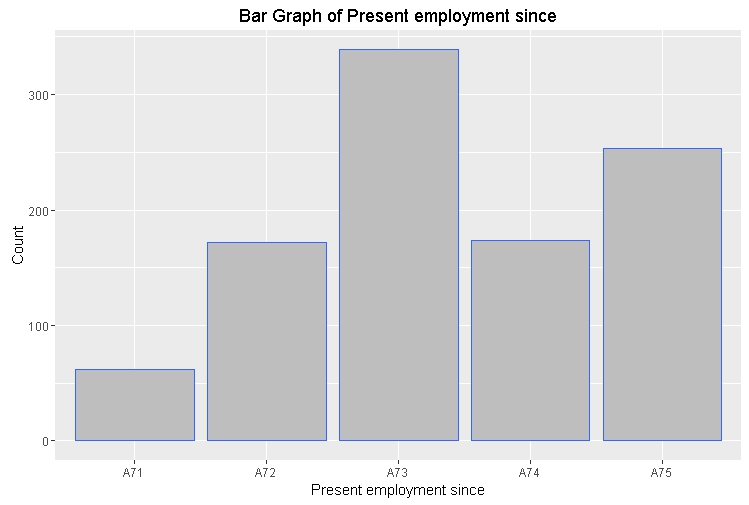
**NOTE:** This should briefly describe the important results and recommendations. The structure is suggestive; make sure to not exceed 7 pages**.**

# Checkpoint-1: Data Understanding and Data Exploration

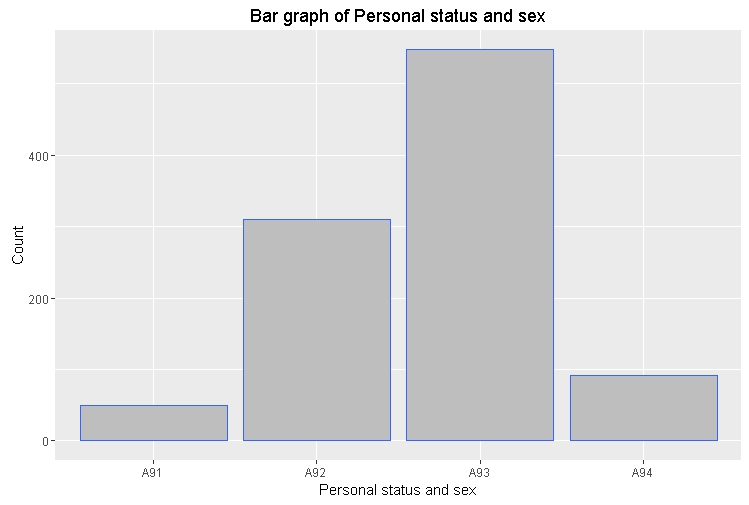
* Display the plots and explain the insights
  + The Below Plot on existing checking account shows that large number of loan is given to people who don't have existing checking account with this bank.
  + The Below Plot on credit history shows that most number of people to whom loan is given has paid their existing credit with this account.



* + The Below Plot on purpose shows that mostly people take credit for radio/television (A43) or for new or old cars (A40, A41).
  + The Below Plot on present employment since shows that most people take loan who are in the current organisation for less than 4 year.



* The Below Plot on Personal status and sex shows that most people who take loan are male and are single.



# 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:
* Explain the methodology of how did you created dummy variables
* If binning for numerical variables done explain why it was required?

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| IQR | Credit. Amount |
| Dummy creation |  |
| Binning of variables | Age.in.Years, Duration.in.month |

# Checkpoint 3: Splitting the Dataset into train and test

* We have divided our final data into 70:30 ration, will have use the sample split function as below :-
* There are 70% TRUE and 30% FALSE values in this vector. We have assigned the TRUE indices as the train data and FALSE indices as the test data.
* The train data set have 700 obs with 57 variables
* The test data sets is of 300 obs and 57 variables

# 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

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeeric)** |
| Credit.amount | 0.4834 |
| `Credit.history : A31` | 0.9401 |
| `Installment.rate.in.percentage.of.disposable.income : 1` | -0.7592 |
| `Purpose : A41` | -1.3809 |
| `Status.of.existing.checking.account : A11` | 1.8873 |
| `Status.of.existing.checking.account : A12` | 1.3660 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 735.9 |
| Null deviance | 855.21 |
| Residual Deviance | 721.90 |

We have use StepAIC to select the model based on AIC value, then we have use VIF to check for any multicollinearity,

Then we have reduce the variable based on its P-Value, we have consider only those variables which have p-value not less than 2. Below is our Model

log(odds) = -1.7153 + 0.4834 (Credit.amount) + 0.9401 (`Credit.history : A31`) - 0.7592 (`Installment.rate.in.percentage.of.disposable.income : 1`) - 1.3809 (`Purpose : A41`) + 1.8873 (`Status.of.existing.checking.account : A11`) + 1.3660 (`Status.of.existing.checking.account : A12`)

Since, from the above model it is clear that an increase in the purpose of loan which is car will decrease its probability to get loan, also an increase in the checking account A12 which is DM > 200, will increase its probability to get loan, so our model is as per the business problem.

# 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 | 7.644655e-01 | C-statistic | 7.180423e-01 |
| KS-statistic | 0.4190476 | KS-statistic | 0.3746032 |
| Model Evaluation (write Accept or Reject) | | Accept | |

We see that the train and test data are very close so our model is working as desire, we also see that the KS- KS-statistic falls under 4th decile in both train and test data, So our model is working good for both train and test data sets.

# Checkpoint 6: Threshold value

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

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 70% |
| Sensitivity | 0.6714 |
| Specificity | 0.7667 |

We have taken the threshold value of .33 since we have to maximum the number of accurately predicting the bad customers.