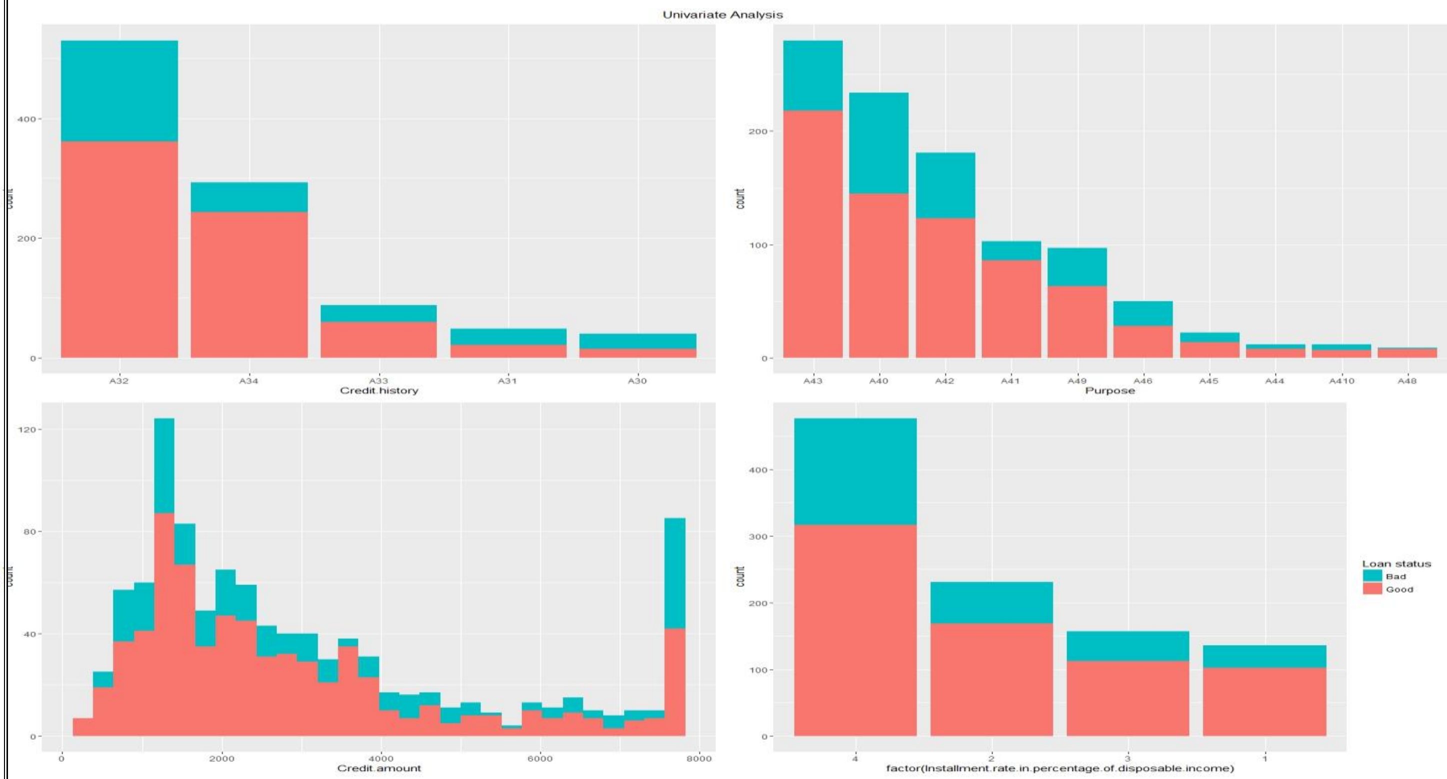
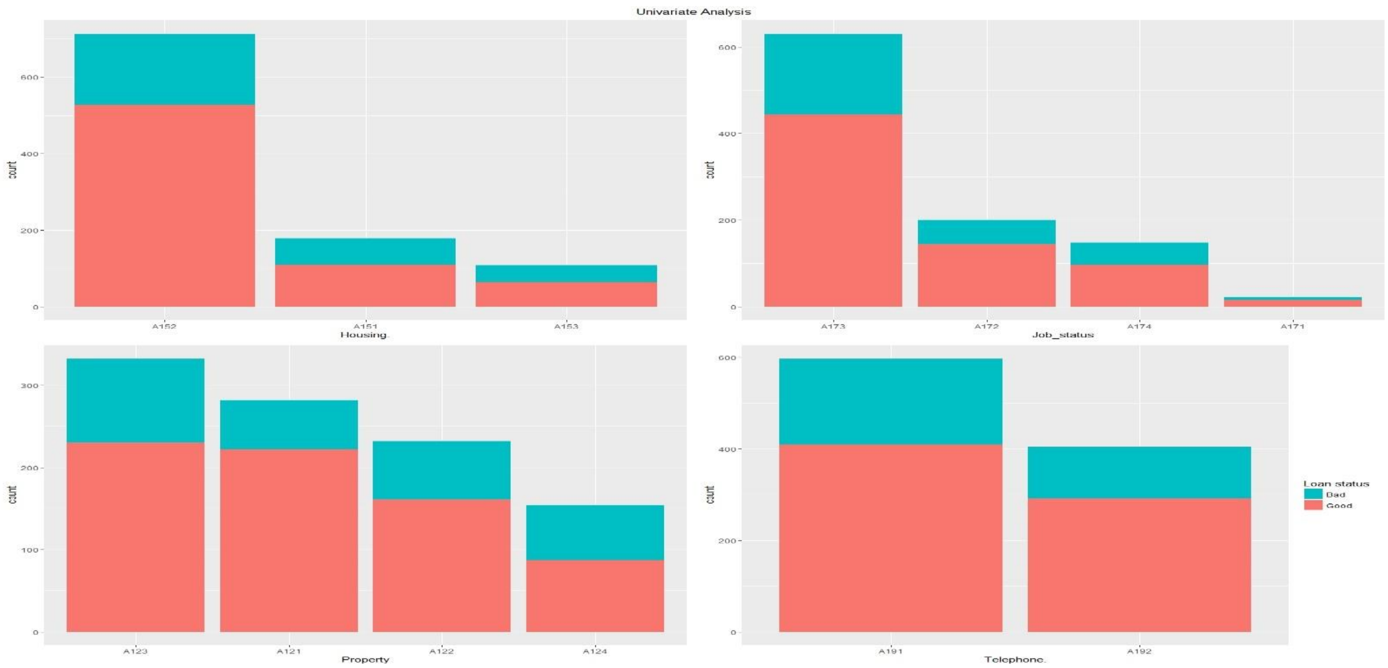
LOGISTIC REGRESSION SUBMISSION

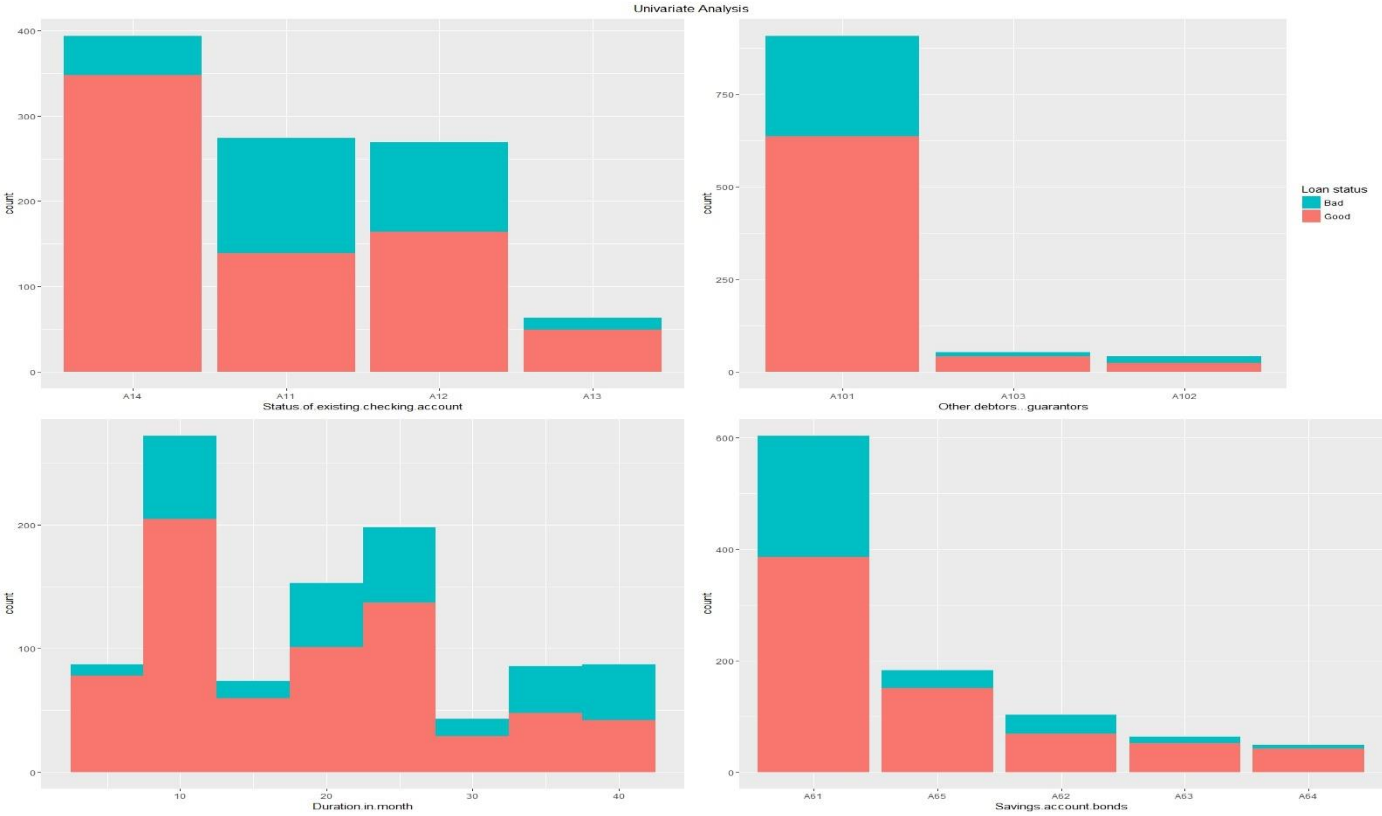
NOTE: This should briefly describe the important results and recommendations. The structure is suggestive; make sure to not exceed 7 pages.

Checkpoint-1: Data Understanding and Data Exploration

- Display the plots and explain the insights







Insights:

Attribute	**Value**	**Number of Defaults**
Job	A173 – Skilled employee / Official	High
Debtors	A101 – No Guarantors	High
Housing	A152 – Own house	High
Checking Account	A11, A12 (DM less than 0 and 200)	High
Credit History	A32 (Existing credits paid back till now) A34 (Other credits existing at other banks)	High
Property (loan)	A123 (car or other)	High
Purpose	A40 (New car) A42 (Furniture or equipment)	High
Duration of loan	10 months	High
Telephone	A191 – No telephone number	High
Savings bond	A-161 (DM less than 100)	High
Instalment Rate	Instalment rate of 4% of disposable income	High
Credit Amount	For credits above 7500 and for credits between 1000 and 2000	High

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 (Including Response variable)
Total missing values in the dataset	NIL

- Explain the methodology of Outlier treatment and fill the below table: The presence of outliers are checked with the help of boxplot. Stats() function. Additionally, a boxplot was plotted to check the presence of outliers. Any outliers found was capped and floored by the nearest quantile which is not considered as an outlier.
- Explain the methodology of how did you created dummy variables –. All the character variables (non-numeric) was selected and created a separate data frame. These variables were converted into factors. Then Dummies package was utilised for creating dummy variables. To reduce the number of dummy variables to one less than the number of factors some dummy columns were dropped. These dummy variables were finally merged with the main data frame. The character variables in the original dataframe was dropped to ensure that only numeric variables are left for modelling.
- If binning for numerical variables done explain why it was required – Binning was not performed

Additionally, fill the below table:

Operations performed	Variable Name
Outlier treatment	Credit.amount, Duration.in.month, Age.in.Years
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, Job status, Telephone, foreign worker,
Binning of variables	NIL

Checkpoint 3: Splitting the Dataset into train and test

Initially the seed was set for 100 to ensure that results are same whenever the code runs.

To ensure same proportion of Good and bad loan class labels are there in both the test and training set, sample.split() function is used. Training and test datasets are divided into 70:30 proportion.

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

  The methodology of building the model is as follows: -

  (a) Initially a model was built with all the variables

  (b) Then using stepwise method all insignificant variables are removed. significant variables were selected.

  (c) Next variables were checked for multicollinearity. VIF values of variables were checked. Variables with high VIF and are insignificant were removed. If the VIF values are all less than the required threshold than their P-values were checked. Variables with high P-values and VIF are selected for removal. Every time an iteration is done AIC values are checked to see if there is any abnormal increase.

  (d) After nine iterations if a variable is removed there is a significant increase in AIC values and also all the variables are significant. As required Hence iterations are stopped and the model is considered as final.

Additionally, fill the below table: This table includes interpretation of co-efficient and business sense

Significant variables in final model (add more rows if requires)	Coefficients value (Numeric)	Change in Log odds when the variable is increased by 1 when everything else is constant	Business Sense
Duration.in.month	0.04675517	Increase in duration by one month will increase the log odds by 0.04675517	Yes. As increase in loan duration is increasing the odds of default
Housing.A152	-0.46786700	Increase in loan for own house by one will decrease the log odds by -0.46786700	Yes. Increase in loan for house owners will decrease the odds of default
Other.installment.plansA143	-0.51102957	Increase in other installment plans(none) by one will decrease the log odds by -0.51102957	No. Giving loan to someone without any instalment plan is reducing the odds of default
Other.debtors...guarantorsA103	-0.90005597	Increase in loan with guarantor by one will decrease the log odds by -0.90005597	Yes. Increase in loan with a guarantor is decreasing the odds of default

Savings.account.bondsA65	-0.55958746	Increase in loan for unknown or no savings account by one will decrease the log odds by -0.55958746	No. Increase in loan for no savings account is decreasing the odds of default
Savings.account.bondsA64	-1.19577162	Increase in loan for DM greater than or equal to 1000 by one will decrease the log odds by -1.19577162	Yes. Increase in loan for DM is decreasing the odds of default
Present.employment.since.A74	-0.62115160	Increase in loan for employment status A74 by one will decrease the log odds by -0.62115160	Yes. Providing the loan to employed (4 to 7 years is decreasing the odds of default.
PurposeA41	-0.93090579	Increase in loan for used car by one will decrease the log odds by -0.93090579	Yes. Providing the loan for loan for used car is decreasing the odds of default.
Credit.historyA34	-1.63892097	Increase in loan for other credits existing at other banks by one will decrease the log odds by -1.63892097	No. Increase in loans for people already at debt is reducing the odds of default.
Credit.historyA33	-1.31674232	Increase in loan for delay in paying back in the past by one will decrease the log odds by -1.31674232	No. Giving loans to people who delayed in payback earlier is reducing the odds of default
Credit.historyA32	-1.07358284	Increase in loan for existing accounts paid back till now by one will decrease the log odds by -1.07358284	Yes. Increase in loan for who already paid back is reducing the odds of default
Status.of.existing.checking.accountA14	-1.85557323	Increase in loan for no checking account by one will decrease the log odds by -1.85557323	No. Giving loan to someone without checking account is reducing odds of default.
Status.of.existing.checking.accountA13	-1.01663527	Increase in loan for checking account greater than or equal to 200 by one will decrease the log odds by -1.01663527	Yes. Giving loan to someone with checking account is reducing the odds of default
Installment.rate.in.percentage.of.disposable.income	0.23056897	Increase in Instalment rate in percentage of disposable income by one unit will increase the log odds by 0.23056897	Yes. As people disposable income is reduced for paying instalment, defaults are increasing.

Final model metrics	Values (Numeric)
AIC value	686.13
Null deviance	855.21
Residual Deviance	656.13

Checkpoint 5: Model Evaluation

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

  Higher values of c –statistic and KS statistic indicates that the model is very Good. C- Statistic should be close to 1. A C-Statistic more than 0.7 is preferred. As the test dataset C-Statistic is 0.76, the model is considered good. A KS-statistic of more than 0.4 is preferred and should lie in top 5 deciles. As the test KS statistic is more than 0.4 and lie in 4th decile the model is considered effective and accepted.

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	8.161662e-01	C-statistic	7.691534e-01
KS-statistic	0.5156463	KS-statistic	0.4444444
Model Evaluation (write Accept or Reject)		Model is Accepted	

Checkpoint 6: Threshold value

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

  Selection of appropriate threshold value depends on the objective and the business problem we are trying to resolve. In the data dictionary the following instructions is given: -

  "It is worse to class a customer as good when they are bad, than it is to class a customer as bad when they are good".

  It means that it is preferred that sensitivity of the model should be high. As there is a trade-off between sensitivity and specificity, if sensitivity is increased beyond a certain threshold specificity will reduce. This in turn will also reduce the overall accuracy of the model.

  For a Threshold level of 0.042 the results are as follows: -

```
         Reference          (positive – Bad customers(1))
Prediction    0    1
         0   33    0
         1  177   90

       Accuracy : 0.41
       Sensitivity : 1.0000
```

```
      Specificity : 0.1571
```

```
As seen above the false negatives are zero i.e  no bad customers are predicted as good
But this increases the false positives i.e predicting the good customers as bad.
```

If the objective is to increase both specificity and sensitivity than the threshold level of 0.28989295 is suitable. The results for the same are as follows: -

```
          Reference
Prediction   0   1
         0 147  23
         1  63  67

               Accuracy : 0.7133
            Sensitivity : 0.7444
            Specificity : 0.7000
```

Depending on the business objective threshold level may be selected. Here I am going ahead with an objective to maximize sensitivity irrespective of accuracy and specificity. The results are included in the table.

Additionally, fill the below table:

Threshold value	Values (Numeric)
Overall Accuracy	0.41
Sensitivity	1.0000
Specificity	0.1571