Report-German Credit Data Analysis

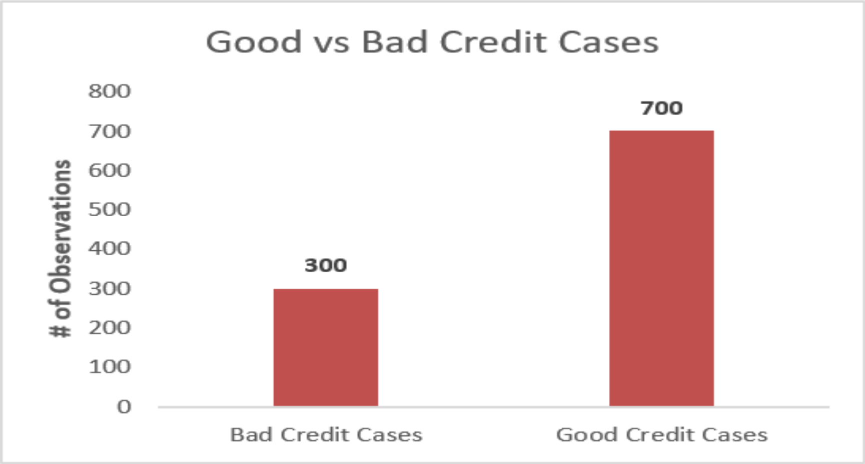
Submitted By-

- **Anupam Sinha**
- **Kush Varma**
- **Vatsal Shah**

Task 1:

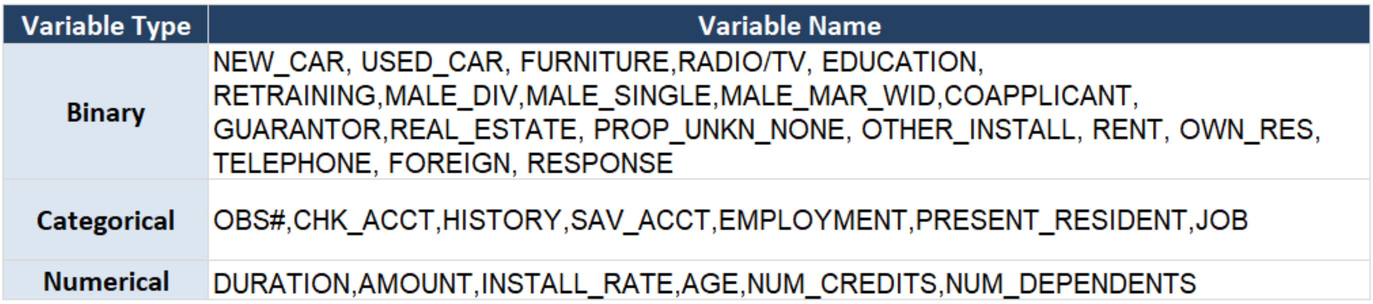
Explore the data: What is the proportion of "Good" to "Bad" cases? Are there any missing values – how do you handle these? Obtain descriptions of the predictor (independent) variables – mean, standard deviations, etc. for real-values attributes, frequencies of different category values. Examine variable plots. Do you notice 'bad' credit cases to be more prevalent in certain value-ranges of specific variables, and is this what one might expect (or is it more of a surprise)?
What are certain interesting variables and relationships (why 'interesting')? From the data exploration, which variables do you think will be most relevant for the outcome of interest, and why?

a) The German Credit Dataset has data on 1000 past credit applicants which has been described by 30 variables. This data contains a "Response" variable which rates the person with a Good or a Bad credit risk. After observing the dataset, the proportion of Good to Bad cases comes to be 7:3 (700:300).



b) Yes, there are few attributes *(majorly binomial variables)* such as New_Car, Used_Car, Furniture, Radio/TV, Education and Retraining which have Null Values. Also, there is an additional quantitative variable "Age" which has null values as well. For the binomial variables, we replaced the missing values by 0 and for the cases where age is null, we replaced those cases by the integer value of mean i.e. 35 *(nine cases)*

c) The dataset consists of 32 attributes and 1000 records. Below is the comprehensive summary of the type of variables which are present in the dataset.



Variable Type	Variable Name
Binary	NEW_CAR, USED_CAR, FURNITURE,RADIO/TV, EDUCATION, RETRAINING,MALE_DIV,MALE_SINGLE,MALE_MAR_WID,COAPPLICANT, GUARANTOR,REAL_ESTATE, PROP_UNKN_NONE, OTHER_INSTALL, RENT, OWN_RES, TELEPHONE, FOREIGN, RESPONSE
Categorical	OBS#,CHK_ACCT,HISTORY,SAV_ACCT,EMPLOYMENT,PRESENT_RESIDENT,JOB
Numerical	DURATION,AMOUNT,INSTALL_RATE,AGE,NUM_CREDITS,NUM_DEPENDENTS

d) Among the 32 variables which are present in the database, below is the summary having the descriptions of the predictor (independent) variables-

Numerical Variables	Minimum	Maximum	Mean	Standard Deviation
DURATION	4	72	20.903	12.052
AMOUNT	250	18,424	3,271.156	2,821.213
INSTALL_RATE	1	4	2.973	1.118
AGE	19	75	35.48	11.31
NUM_CREDITS	1	4	1.407	0.577
NUM_DEPENDENTS	1	2	1.155	0.361

e) Below is the quick summary indicating the frequencies of the categorical and binary variables-

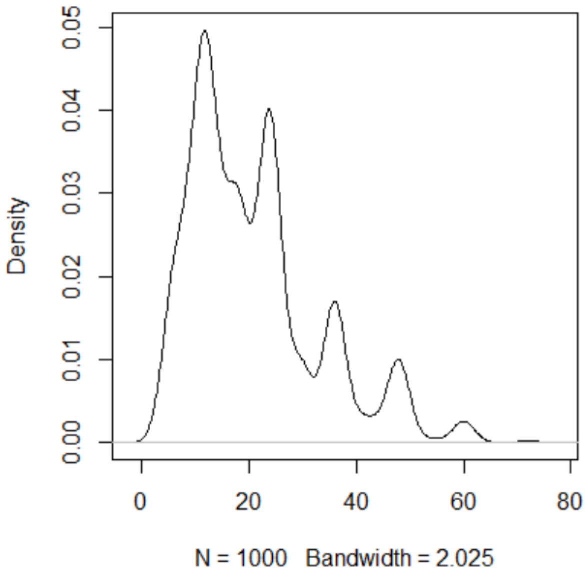
Categorical Variables	0	1	2	3	4
CHK_ACCT	274	269	63	394	NA
HISTORY	40	49	530	88	293
SAV_ACCT	603	103	63	48	183
EMPLOYMENT	62	172	339	174	253
PRESENT_RESIDENT	NA	130	308	149	413
JOB	22	200	630	148	NA

Binary Variables	0	1
NEW_CAR	766	234
USED_CAR	897	103
FURNITURE	819	181
RADIO/TV	720	280
EDUCATION	950	50
RETRAINING	903	97
MALE_DIV	950	50
MALE_SINGLE	452	548
MAR_MAR_WID	908	92

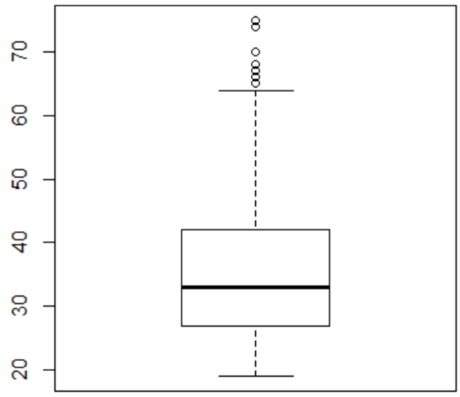
Binary Variables	0	1
COAPPLICANT	959	41
GUARANTOR	948	52
REAL_ESTATE	718	282
PROP_UNKN_NONE	846	154
OTHER_INSTALL	814	186
RENT	821	179
OWN_RES	287	713
TELEPHONE	596	404
FOREIGN	963	37

f) **Variable Plots:** For Univariate Analysis, we have used Bar Plots, Histograms, Density Plots and Boxplots to represent the data.

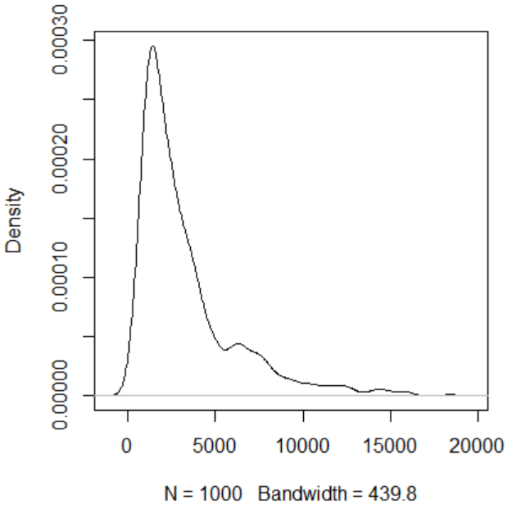
Density Plot to show the Duration:

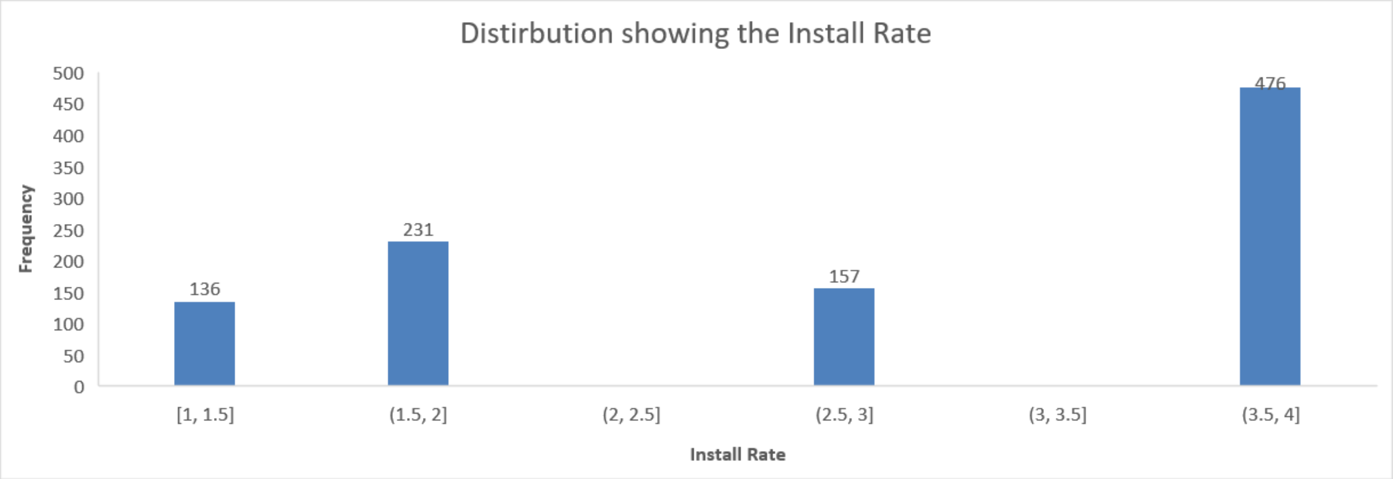


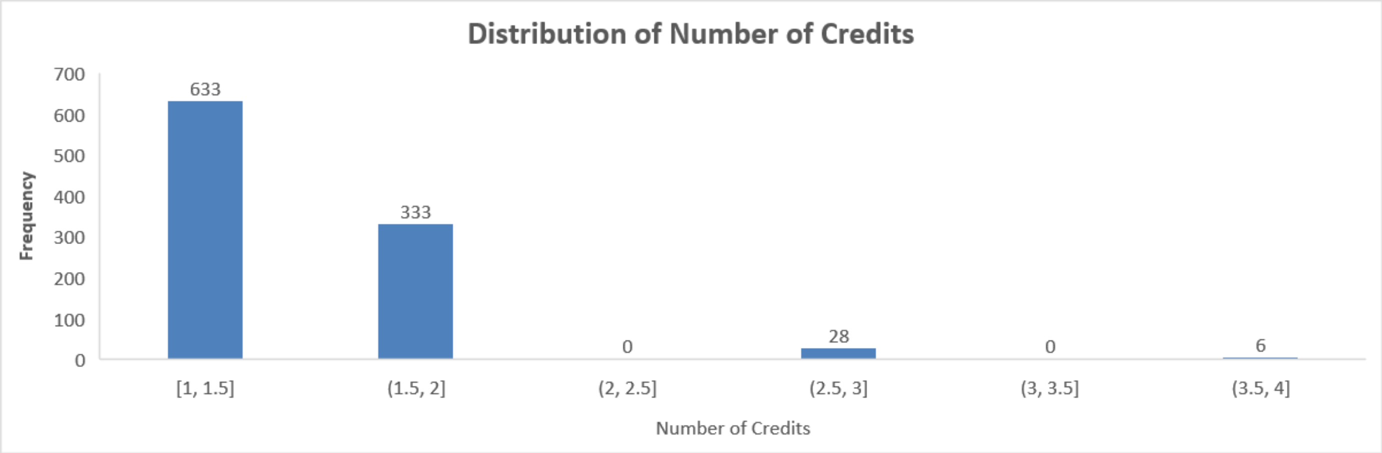
Box Plot to show the Age:

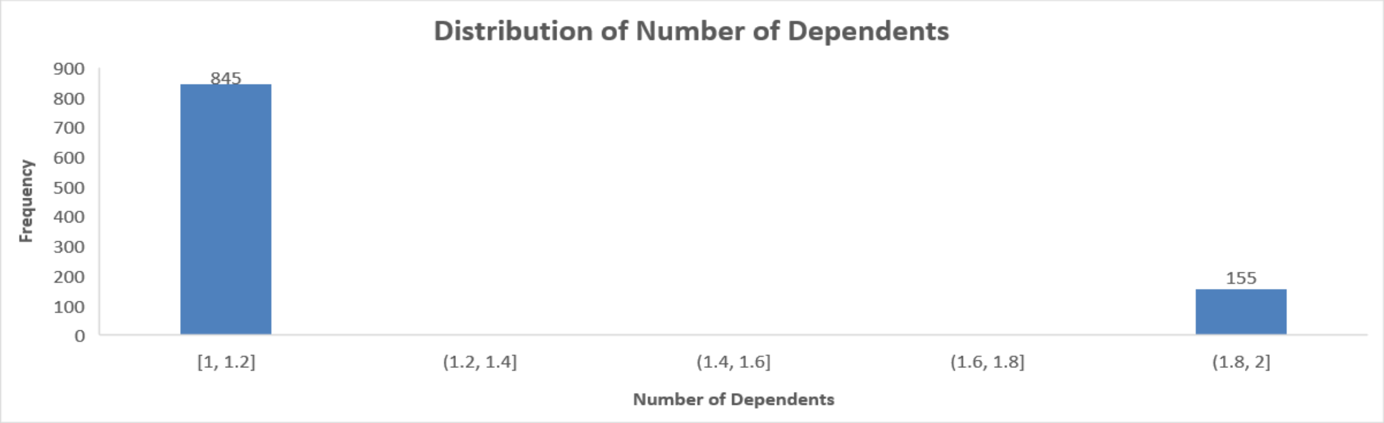


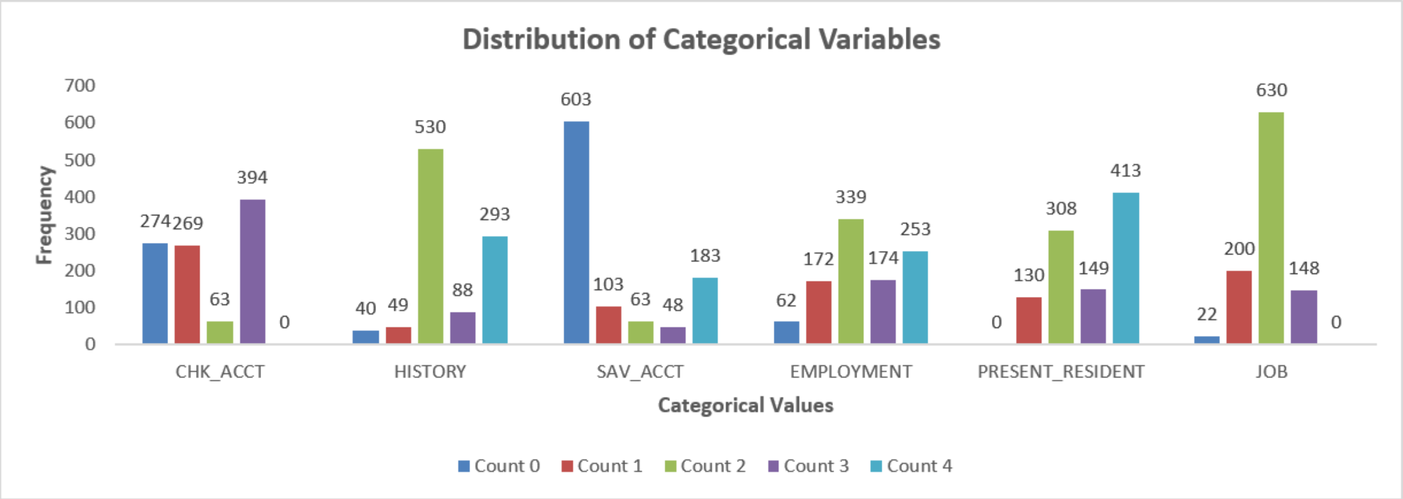
Density Curve to show the Amount:

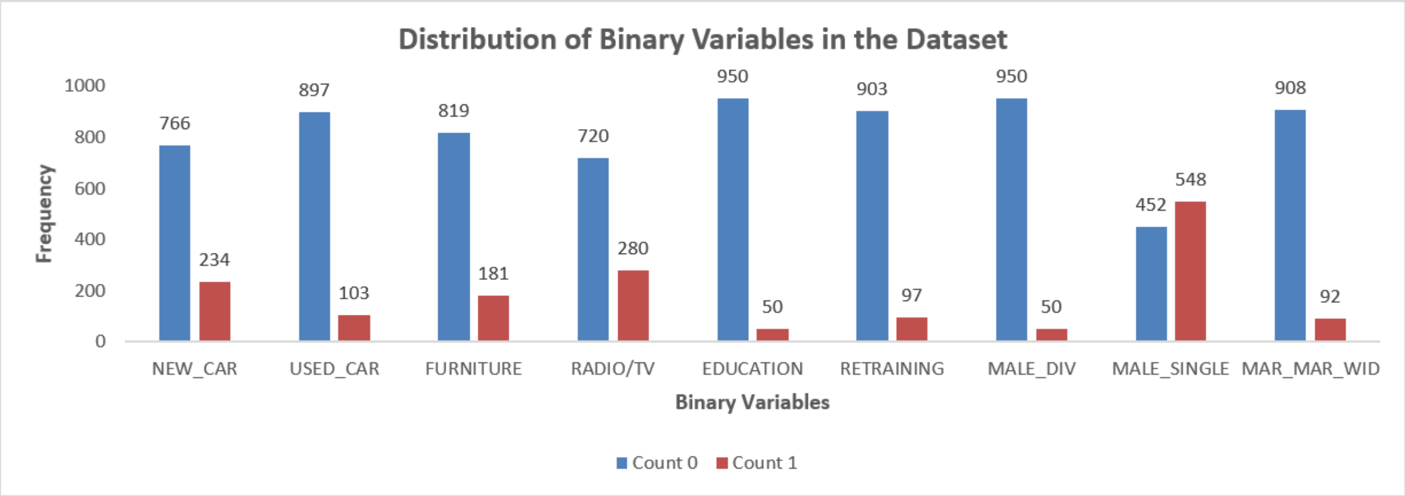


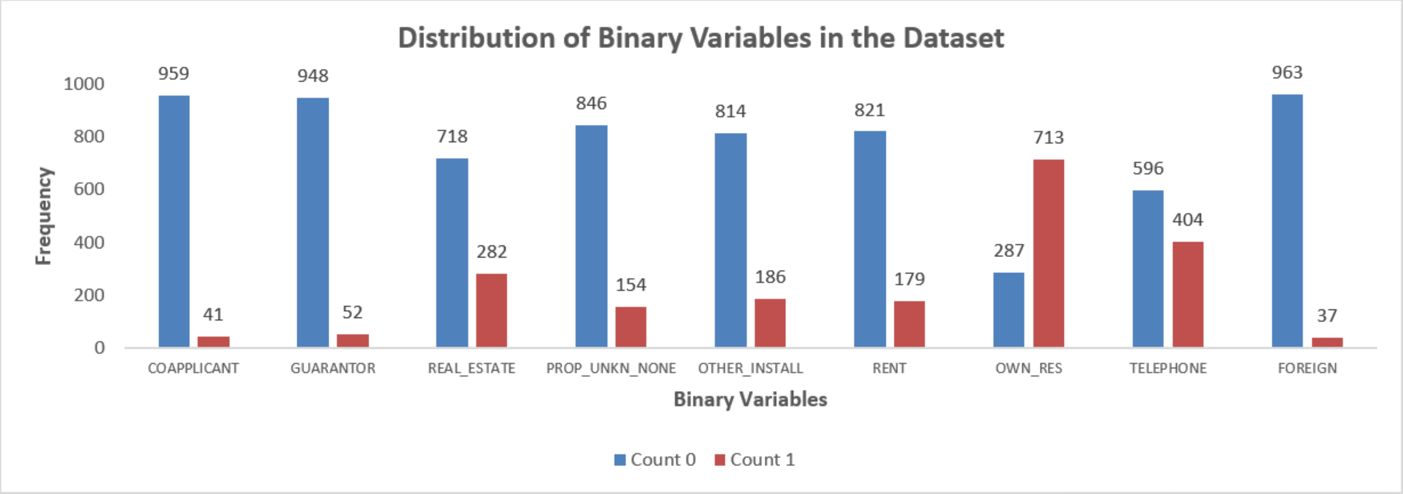

Distirbution showing the Install Rate
Frequency
500
450
400
350
300
250
200
150
100
50
0
136
231
157
476
[1, 1.5]
(1.5, 2]
(2, 2.5]
(2.5, 3]
(3, 3.5]
(3.5, 4]
Install Rate


Distribution of Number of Credits
Frequency
700
600
500
400
300
200
100
0
633
333
0
28
0
6
[1, 1.5]
(1.5, 2]
(2, 2.5]
(2.5, 3]
(3, 3.5]
(3.5, 4]
Number of Credits


Distribution of Number of Dependents
Frequency
900
800
700
600
500
400
300
200
100
0
845
155
[1, 1.2]
(1.2, 1.4]
(1.4, 1.6]
(1.6, 1.8]
(1.8, 2]
Number of Dependents


Distribution of Categorical Variables
Frequency
700
600
500
400
300
200
100
0
274
269
63
394
0
40
49
530
88
293
603
103
63
48
183
62
172
339
174
253
0
130
308
149
413
22
200
630
148
0
CHK_ACCT
HISTORY
SAV_ACCT
EMPLOYMENT
PRESENT_RESIDENT
JOB
Categorical Values
Count 0
Count 1
Count 2
Count 3
Count 4





g) There were couple of cases where we observed bad cases to be prevalent, for instance, when savings account's balance is less than 100 DM we observe the bad credit to be highest. We observe a similar trend in the Checking account balance as well. Also, in case of Duration, bad cases are prevalent the duration is between 12-36 months.

h) Ironically, in the observations stated above for checking account and savings account we observe the good credit cases to be highest in the same category as that of bad cases. Hence, it will be interesting to know the other factors that can help determine the credit response.
While observing the cross tab between the savings account and Employment, we observe ***the most amount of good cases and bad cases in the bucket (2) of Employment which is 1 to 4 years***. Ideally, this should have been more prevalent in the lower buckets as the savings account balance for them is <100 DM

Savings Account vs Employment Rate (Good Credit)

	Employment				
Savings Account	0 : unemployed	1: < 1 year	2 : 1 <= ... < 4 years	3 : 4 <=... < 7 years	4 : >= 7 years
0 : < 100 DM	23	67	140	70	86
1 : 100<= ... < 500 DM	4	7	21	20	17
2 : 500<= ... < 1000 DM	3	2	20	8	19
3 : =>1000 DM	0	6	16	8	12
4 : unknown/ no savings account	9	20	38	29	55

Savings Account vs Employment Rate (Bad Credit)

	Employment				
Savings Account	0 : unemployed	1: < 1 year	2 : 1 <= ... < 4 years	3 : 4 <=... < 7 years	4 : >= 7 years
0 : < 100 DM	17	53	70	30	47
1 : 100<= ... < 500 DM	3	10	12	4	5
2 : 500<= ... < 1000 DM	0	3	6	1	1
3 : =>1000 DM	0	1	2	1	2
4 : unknown/ no savings account	3	3	14	3	9

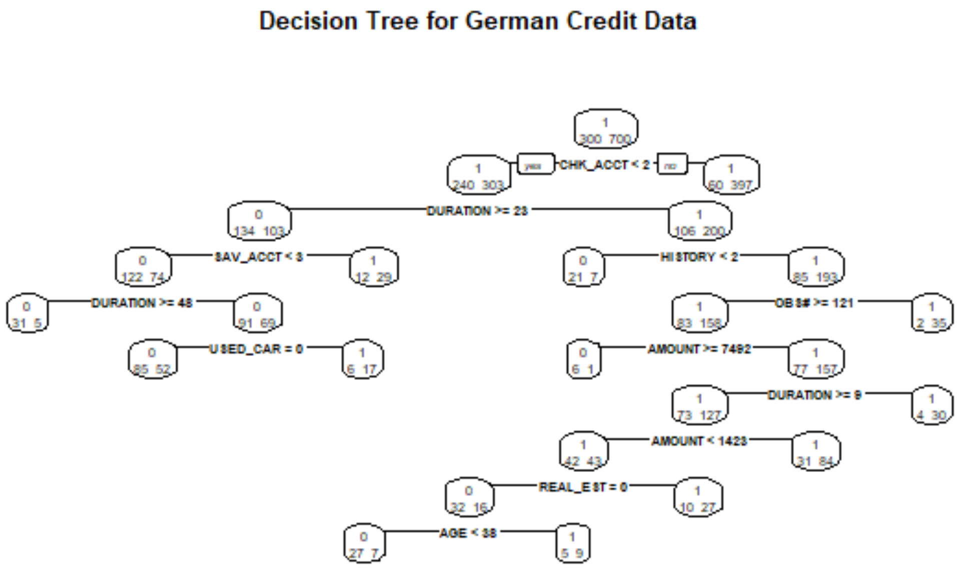
i) Based on our analysis, we believe the following variables will be the most relevant:

1. **Duration:** As the name suggests, this variable states the duration of credit history and ideally longer the duration should specify more credit history available of an individual.
2. **Age:** This variable can help us bifurcate the German Credit Dataset to understand the different segments of applicants based on the same age groups.
3. **History:** This variable helps in determining a person's credit history.
4. **Amount:** This variable gives us the credit amount of each applicant and could be helpful for our decision tree
5. **CHK_ACCAT and SAV_ACCT:** These variables will provide us the details of the account balances of each applicant. variable will help understand the account status of each applicant. We have ~70% of the applicants who have either checking or savings account and it could be our formative audience.

Task-2:

We will first focus on a descriptive model – i.e. assume we are not interested in prediction. (a) Develop a decision tree on the full data (using the rpart package).
What decision tree node parameters do you use to get a good model. Explain the parameters you use.
(b)Which variables are important to differentiate "good" from "bad" cases – and how do you determine these? Does this match your expectations (from your response in Question 1)?
(c)What levels of accuracy/error are obtained? What is the accuracy on the "good" and "bad" cases? Obtain and interpret the lift chart. Do you think this is a reliable (robust?) description, and why.

a) Decision Tree using the Full Data


Decision Tree for German Credit Data

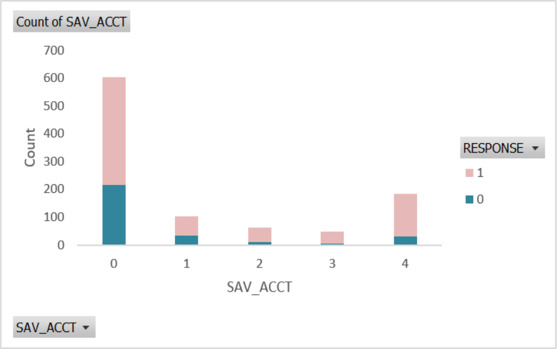
What decision tree node parameters do you use to get a good model. Explain the parameters you use

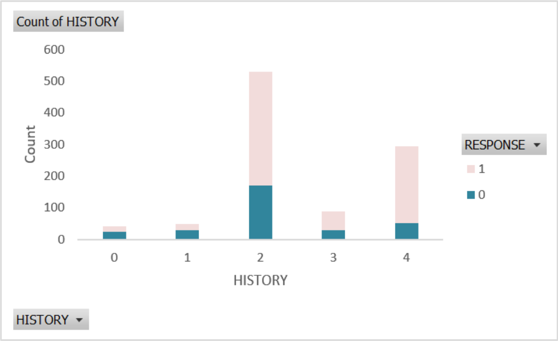
Split type	Parameter	Accuracy
Full Data	Information	82.2
	Information Minsplit= 30, Minibucket=10, cp=0	79.2
	Gini	83

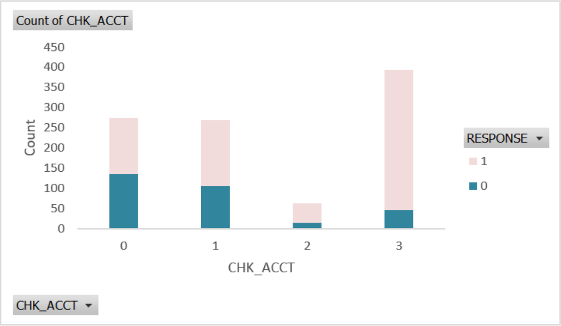
Applying the node parameter as above gives an accuracy of more than 80% for some parameters which is an indication of over-fitting of the model in the data. We thus choose **Gain Ratio** to be the node parameter in the decision tree which gives us an average accuracy of ~79 %. Information Gain Ratio reduces bias greatly and thus produces the best suited model.

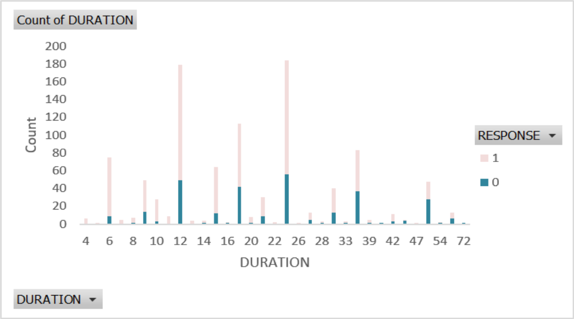
b) Which variables are important to differentiate "good" from "bad" cases – and how do you determine these? Does this match your expectations (from your response in Question 1)?

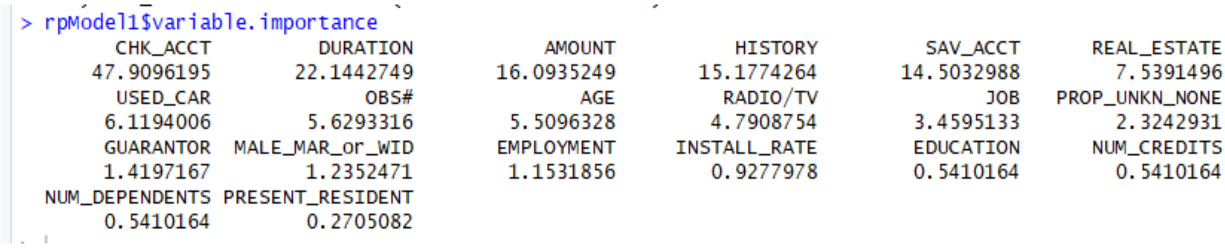
The important variables to differentiate good from bad cases are ***CHK_ACCT, DURATION, AMOUNT, HISTORY, SAV_ACCT***. This can be examined from the following output:











```
> rpModel1$variable.importance
       CHK_ACCT        DURATION          AMOUNT         HISTORY        SAV_ACCT     REAL_ESTATE
     47.9096195      22.1442749      16.0935249      15.1774264      14.5032988       7.5391496
       USED_CAR            OBS#             AGE        RADIO/TV             JOB  PROP_UNKN_NONE
      6.1194006       5.6293316       5.5096328       4.7908754       3.4595133       2.3242931
      GUARANTOR  MALE_MAR_or_WID      EMPLOYMENT    INSTALL_RATE       EDUCATION     NUM_CREDITS
      1.4197167       1.2352471       1.1531856       0.9277978       0.5410164       0.5410164
 NUM_DEPENDENTS PRESENT_RESIDENT
      0.5410164       0.2705082
```

After the initial analysis, we expected ***CHK_ACCT, DURATION, AMOUNT, HISTORY, SAV_ACCT*** to have higher variable importance and this can be justified to an extent using the numerical values stated above. These have been determined using Information Gain ratio. The ratio is found to be high for above mentioned variables

c) What levels of accuracy/error are obtained? What is the accuracy on the "good" and "bad" cases? Obtain and interpret the lift chart.

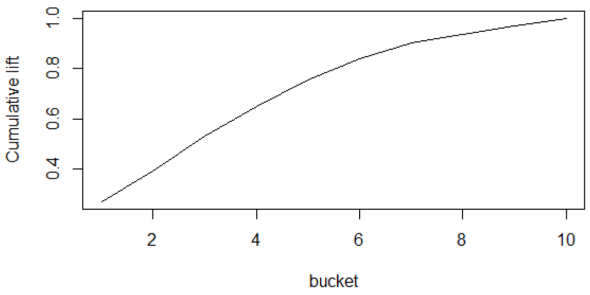
➔ The mean accuracy which we observe from the above decision tree is 79.8%. Below is the Confusion matrix which has been used to compute the accuracy on the good and bad cases-

	True	
Prediction	0	1
0	170	72
1	130	628

➔ **Accuracy of good cases:** ((170+628)/1000) *100 = 79.8%

➔ **Accuracy of Bad Cases:** ((72+130)/1000) *100 = 20.2%

Lift Chart:



Interpretation of Lift Chart:

Using certain explanatory variables, the AUC for the dataset is 0.778. Considering all the variables, there is a gradual Improvement in AUC value to 0.798. Likewise, if many trees are aggregated along with their predictions, the model will perform better like the Random forest model

Do you think this is a reliable (robust?) description, and why.

A model is said to be reliable if the dataset is divided into training and test data to draw results. In this case, since entire data has been used for data modelling, the final estimated tree would be influenced by minute change in data and hence not robust

Task- 3:

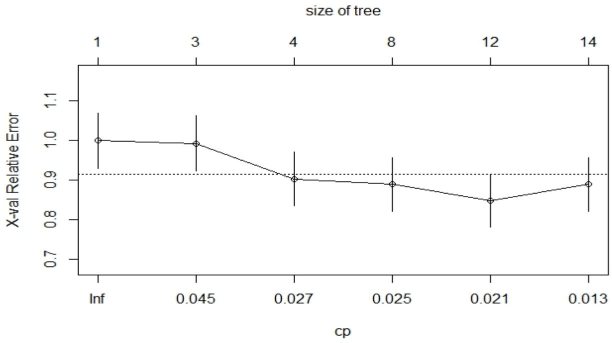
a. **Develop decision trees using the rpart package. What model performance do you obtain? Consider performance based on overall accuracy/error and on the 'good' and 'bad' credit cases – explain which performance measures, like recall, precision, sensitivity, etc. you use and why. Also consider lift, ROC and AUC.**
In developing the models above, change decision tree options as you find reasonable (for example, complexity parameter (cp), the minimum number of cases for split and at a leaf node, the split criteria, etc.) - explain which parameters you experiment with and why. Report on if and how different parameters affect performance. Which decision tree parameter values do you find to be useful for developing a good model?
Describe the pruning method used here. How do you examine the effect of different values of? cp, and how do you select the best pruned tree?

We developed the simple/default decision tree through the rpart package. Since the model was one of the most primitive ones we got an accuracy as high as 83.4%. The good cases had an accuracy of higher than the bad cases.

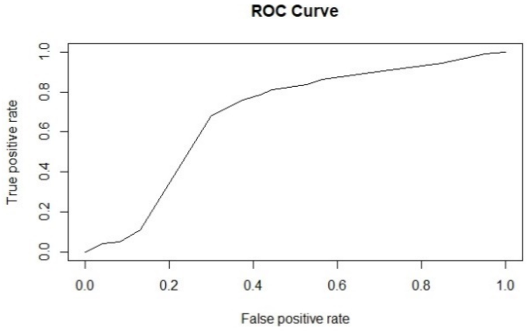
The different performance measures like recall, precision and sensitivity came out to be much higher in the case of 1s rather than the 0s.

```
precision
        0         1
 0.7807018 0.8264249
recall
        0         1
 0.5705128 0.9273256
f1
        0         1
 0.6592593 0.8739726
```

So, if we use this model in making sure we have 1s to deal with, our model is great. But if it comes to using them where 0s have a greater weightage, we might need to think once more



Since all the 1s have greater performance, our ROC curve and subsequently our AUC is coming good.



AUC comes 68.6% which looks fair enough. We use other values of cp, maxdepth, Minsplit to find if there's a difference in any of those. Here's a table defining all of that.

	Accuracy	Precision	Recall	FPRate
Model (Gini)- Default	0.716	0.585	0.346	0.235
Cp=0.001	0.719	0.591	0.358	0.230
Minsplit = 20	0.712	0.538	0.379	0.232
Maxdepth = 15	0.720	0.572	0.347	0.239
Model (Information Gain) - Default	0.714	0.571	0.380	0.220
Cp=0.001	0.717	0.575	0.345	0.227
Minsplit = 20	0.698	0.580	0.362	0.218
Maxdepth = 15	0.713	0.568	0.340	0.223

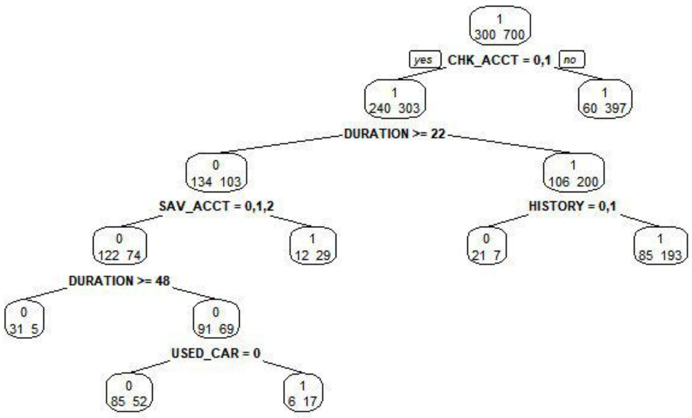
Here we found that the Gini model with maxdepth of 15 was giving us high accuracy and for now we are going ahead with choosing that model. Maxdepth and Minsplit both are very good measures to make sure all the minute details are taken care off.

According to the FPrate, almost 22% of the "bad" creditors are being recognized as "good".

The **pruning method** used here is the complexity parameter pruning. CP is the complexity parameter to where the decision tree splits. It effects the accuracy to a great extent and precision. Sensitivity and specificity are also enhanced due to this pruning method.

b. **Consider another type of decision tree – C5.0 – experiment with the parameters till you get a ‘good’ model. Summarize the parameters and performance you obtain.**

This is how one of our c5.0 models look like:



The measure of the performance parameters went a little bit up with this kind of c5.0 model. But when compared with ROC and AUC, the model significantly lost foot and thus, we won’t be using it as the best fit model going ahead.

c. **Decision tree models are referred to as ‘unstable’ – in the sense that small differences in training data can give very different models. Examine the models and performance for different samples of the training/test data (by changing the random seed). Do you find your models to be unstable -- explain?**

While changing seed value we came across some questionable routes. We found that the seed change will bring startling changes in the performance matrices but will not make significant change in the accuracy. Here’s the table for the different kind of seeds we used and the accuracy of each. (The model we used for this is the best for us till now – Gini and with cp-0.001 with Minsplit=20 and maxdepth =15)

	Seed (123)	Seed (5)	Seed (400)	Seed (1300)
Accuracy	0.723	0.729	0.719	0.689

We see they are unstable and as we go and increase the seed value they have a tendency to decrease and thus we’ll be going with the default of seed -123.

d. **Which variables are important for separating ‘Good’ from ‘Bad’ credit? Determine variable importance from the different ‘best’ trees. Are there similarities, differences?**

The best tress in different gave us the variable importance in different manner. But for most of the cases the following are the variable that are used for separating the ‘good’ and ‘bad’ creditors. They are:

1) CHK_ACCT
2) DURATION
3) AMOUNT
4) HISTORY

There are different models which gives us other rankings too. But the above mentioned 4 are common in all and are the foremost as well. They are not similar because trees in different models take a different route and have different results.

e. **Consider partitions of the data into 70% for Training and 30% for Test, and 80% for Training and 20% for Test and report on model and performance comparisons.**

Out of 50:50 Split, 70:30 Split and 80:20 Split, the best model is 70:30 split. The accuracy for 80:20(71.2) split is slightly less than 70:30(72.1) split in the test sets.
The 70-30 split also gives us an accuracy higher in the training sets, 79% to 76% in the 80-20 split.
Thus, we are going with the 70-30 as the best split out the 3.

We shall be ***preferring the 70-30 split*** and the model with maxdepth=15 and Minsplit=20 and cp value =0.001.

Task- 4:

Use the misclassification costs to assess performance of a chosen model from Q 2 above. Compare model performance. Examine how different cutoff values for classification threshold make a difference. Use 3 the ROC curve to choose a classification threshold which you think will be better than the default 0.5. What is the best performance you find?
(b) Calculate and apply the 'theoretical' threshold and assess performance – what do you notice, and how does this relate to the answer from (a) above.
(c) Use misclassification costs in building the tree models (rpart and C5.0) – are the trees here different than ones obtained earlier? Compare performance of these two new models with those obtained earlier (in part 3a, b above).

a) For a 70:30 scenario obtained from the ROC curve above, the model performance for different thresholds can be assessed as under:

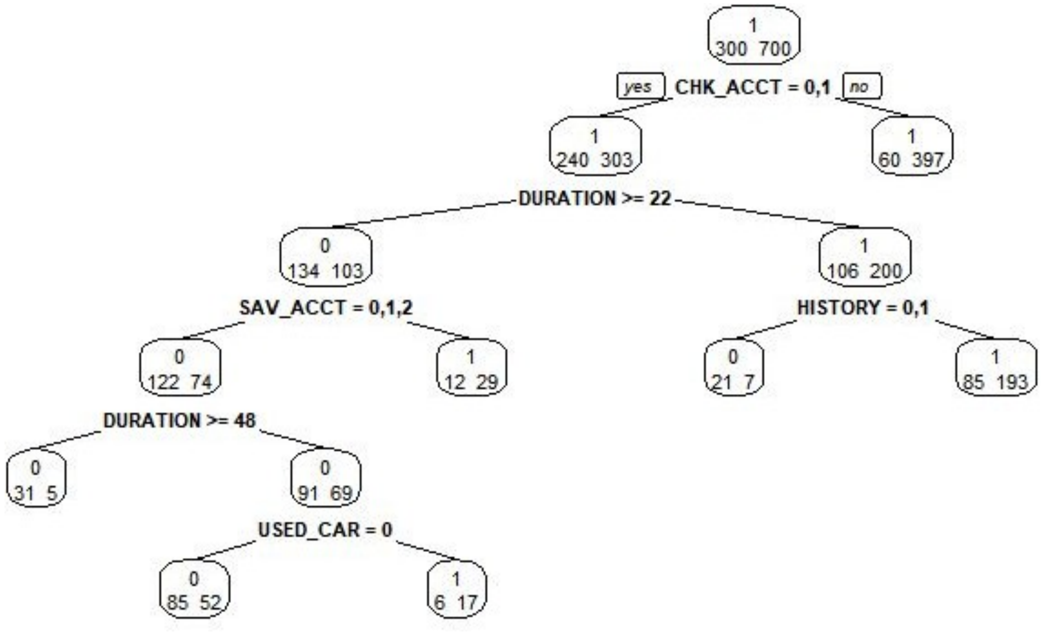
	Misclassification Cost		Performance	
Threshold	Training (70%)	Testing (30%)	Training (70%)	Testing (30%)
0.5	73.8	68.2	66.8	66.2
0.6	71.2	69.8	69.3	72.1
0.7	67.1	68.1	78.4	75.7
0.8	57.1	69.2	63.9	62.5

We find the best performance with threshold 0.7 as the performance of training and test dataset is above 70%. For other threshold, although the cost is less, it compensates with low accuracy as well.

b) Applying the theoretical threshold of 0.83

true/ pred	**0**	**1**
0	168	193
1	38	301
Mean Accuracy: 0.67		

C-50 Decision Tree:



As per the analysis, the most appropriate case would be with threshold value of 0.7,as the performance of the training and test dataset in this above is 70%. For other thresholds the cost is low but the accuracy is also less.

Task-5:

Let’s examine your ‘best’ decision tree model obtained. What is the tree depth? And how many nodes does it have? What are the important variables for classifying “Good’ vs ‘Bad’ credit?
Identify two relatively pure leaf nodes. What are the ‘probabilities for ‘Good’ and ‘Bad’ in these nodes?

→The ‘best decision tree model according to our findings is the ***‘gini’ model with a 70-30 split*** on the training and the test data. Max depth = 15. Minsplit = 20. Cp = 0.001
The important variables for classifying are:
CHK_ACCT
AMOUNT
SAV_ACCT
DURATION

Two relatively pure nodes are as follows:

- #144 AMOUNT< 1539.5 13 1 0 (0.92307692 0.07692308)

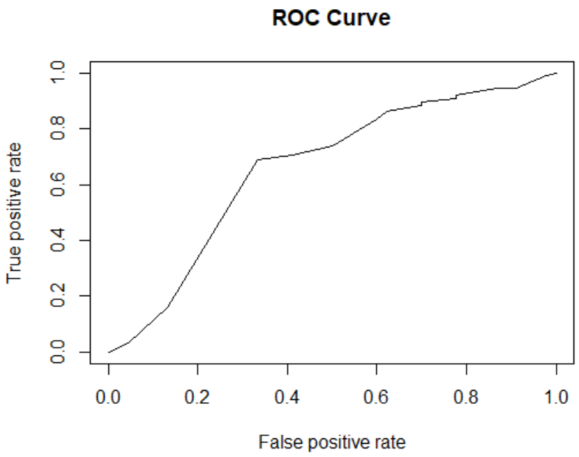
  Node #144 which says if the AMOUNT is less than 1539.5, then there is a 92.307% probability that there will be a 1.
- #5 SAV_ACCT>=3.5 20 2 1 (0.10000000 0.90000000)

  Node #5 which says if the SAV_ACCT are greater than or equal to 3.5, there is a 90% probability that there will be a 0.

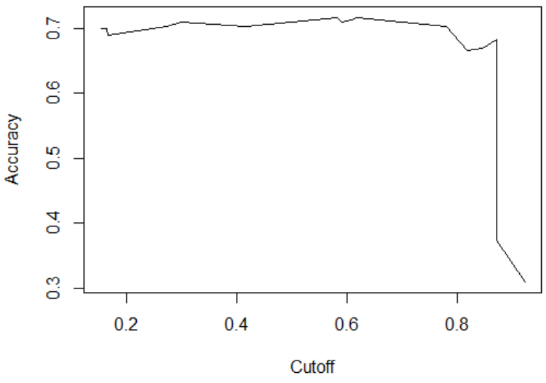
Task-6:

a) **The predicted probabilities can be used to determine how the model may be implemented. We can sort the data from high to low on predicted probability of “good” credit risk. Then, going down the cases from high to low probabilities, one may be able to determine an appropriate cutoff probability - values above this can be considered acceptable credit risk. The use of cost figures given above can help in this analysis.**

For this, first sort the validation data on predicted probability. Then, for each validation case, calculate the actual cost/benefit of extending credit. *The cost has come out to be 0.136.*
Add a separate column for the cumulative net cost/benefit. *The max profit has been calculated which is 0.70*. **How far into the validation data would you go to get maximum net benefit? In using this model to score future credit applicants, what cutoff value for predicted probability would you recommend?**
As shown below in the ROC Curve the cutoff is known to be 0.87



AUC CURVE


Accuracy
Cutoff