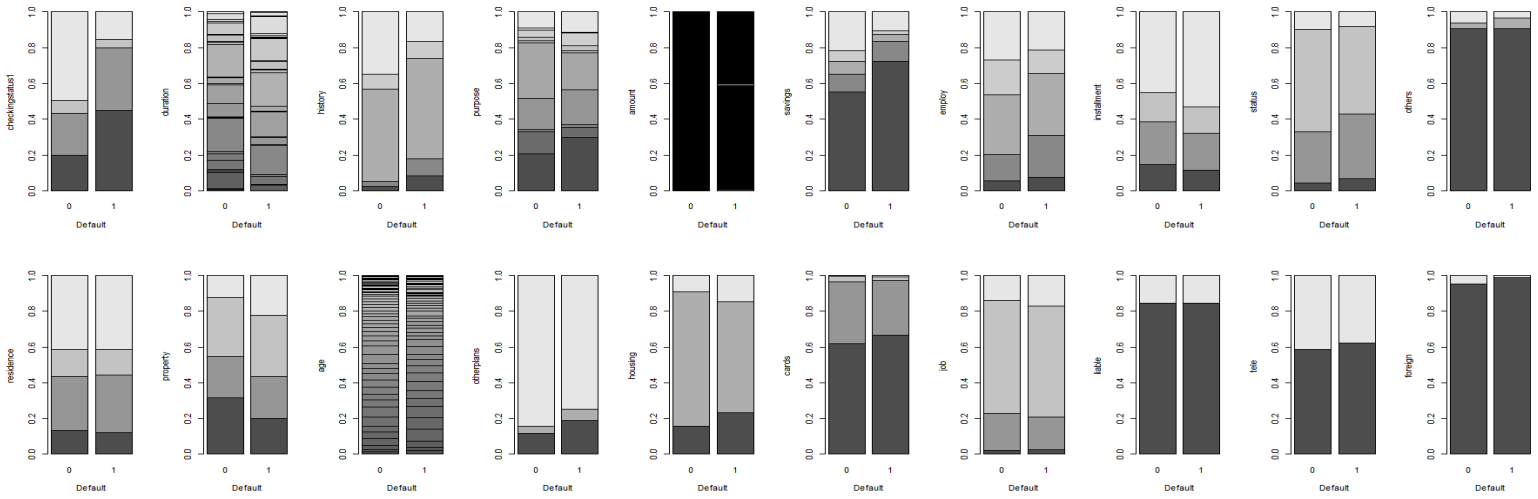
Problem 1 Section 1

a) German credit data set contains 13 categorical and 7 numerical prodictors for 1 response variable named Default (binary) with 1000 observations.

```
'data.frame':  1000 obs. of  21 variables:
 $ Default        : int  0 1 0 0 1 0 0 0 0 1 ...
 $ checkingstatus1: Factor w/ 4 levels "A11","A12","A13",..: 1 2 4 1 1 4 4 2 4 2 ...
 $ duration       : int  6 48 12 42 24 36 24 36 12 30 ...
 $ history        : Factor w/ 5 levels "A30","A31","A32",..: 5 3 5 3 4 3 3 3 3 5 ...
 $ purpose        : Factor w/ 10 levels "A40","A41","A410",..: 5 5 8 4 1 8 4 2 5 1 ...
 $ amount         : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
 $ savings        : Factor w/ 5 levels "A61","A62","A63",..: 5 1 1 1 1 5 3 1 4 1 ...
 $ employ         : Factor w/ 5 levels "A71","A72","A73",..: 5 3 4 4 3 3 5 3 4 1 ...
 $ installment    : int  4 2 2 2 3 2 3 2 2 4 ...
 $ status         : Factor w/ 4 levels "A91","A92","A93",..: 3 2 3 3 3 3 3 3 1 4 ...
 $ others         : Factor w/ 3 levels "A101","A102",..: 1 1 1 3 1 1 1 1 1 1 ...
 $ residence      : int  4 2 3 4 4 4 4 2 4 2 ...
 $ property       : Factor w/ 4 levels "A121","A122",..: 1 1 1 2 4 4 2 3 1 3 ...
 $ age            : int  67 22 49 45 53 35 53 35 61 28 ...
 $ otherplans     : Factor w/ 3 levels "A141","A142",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ housing        : Factor w/ 3 levels "A151","A152",..: 2 2 2 3 3 3 2 1 2 2 ...
 $ cards          : int  2 1 1 1 2 1 1 1 1 2 ...
 $ job            : Factor w/ 4 levels "A171","A172",..: 3 3 2 3 3 2 3 4 2 4 ...
 $ liable         : int  1 1 2 2 2 2 1 1 1 1 ...
 $ tele           : Factor w/ 2 levels "A191","A192": 2 1 1 1 1 2 1 2 1 1 ...
 $ foreign        : Factor w/ 2 levels "A201","A202": 1 1 1 1 1 1 1 1 1 1 ...
```

b) To select the significant predictors, fit `glm` for each predictor and check the chi square value from `anova` table. 0.25 was the cut off value for selecting the significant predictors.



	respond	Chi-value	decision
1	checkingstatus1	2.787203e-28	ueseful
2	duration	2.398744e-11	ueseful
3	history	2.313958e-12	ueseful
4	purpose	7.268797e-05	ueseful
5	amount	1.928791e-06	ueseful
6	savings	7.049052e-08	ueseful
7	employ	1.146430e-03	ueseful
8	installment	2.116372e-02	ueseful
9	status	2.396272e-02	ueseful
10	others	3.597102e-02	ueseful
11	residence	9.252355e-01	unuseful
12	property	3.106302e-05	ueseful
13	age	3.389931e-03	ueseful
14	otherplans	2.129794e-03	ueseful
15	housing	1.448466e-04	ueseful
16	cards	1.443321e-01	ueseful
17	job	6.032614e-01	unuseful
18	liable	9.239826e-01	unuseful

```
19            tele 2.477554e-01  ueseful
20         foreign 4.494505e-03  ueseful
```

Bellow shows the bar plots of each predictor for the Default variable.

Thus the predictors `residence, job and liable` were insignificant. A new `glm` fit was done with the remaining predictors and it was compared with the full model.

```
> anova(fit.full, fit2, test = "Chisq")
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       956     897.69
2       951     895.82  5   1.8675   0.8672
```

Based on the `anova` table, we can reject the full model.

Next, by checking the p-values on the summary for the fits, insignificant predictors removed from the model one by one and compared the new model with the full model using chi square value.

`employ` is insignificant. Fit a new model (`fit3`)

```
> anova(fit.full, fit3, test = "Chisq")
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       951     895.82
2       960     905.41 -9  -9.5971   0.3841
```

Reject the full model.

`property` is insignificant. Fit a new model (`fit4`)

```
> anova(fit.full, fit4, test = "Chisq")
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       951     895.82
2       963     908.82 -12  -13.001    0.369
```

Reject the full model.

`age` is insignificant. Fit a new model (`fit5`)

```
> anova(fit.full, fit5, test = "Chisq")
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       951     895.82
2       964     911.54 -13  -15.724   0.2643
```

Reject the full model.

`cards` is insignificant. Fit a new model (`fit6`)

```
> anova(fit.full, fit6, test = "Chisq")
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       951     895.82
2       965     912.96 -14  -17.147   0.2484
```

Reject the full model.

`tele` is insignificant. Fit a new model (`fit7`)

```
> anova(fit.full, fit7, test = "Chisq")
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       951     895.82
2       966     916.03 -15  -20.214   0.1639
```

Reject the full model.

This gives the best-fitted model.

c)

$$\log\left(\frac{Default}{1-Default}\right) = 1.235 + \begin{cases}(-3.861\text{e}-01)*\text{A12}\\(-1.032\text{e}+00)*\text{A13}\\(-1.737\text{e}+00)*\text{A14}\end{cases} + (2.838\text{e}-02)*\text{duration} + \begin{cases}(-1.314\text{e}-01)*\text{A31}\\(-8.379\text{e}-01)*\text{A32}\\(-9.662\text{e}-01)*\text{A33}\\(-1.545\text{e}+00)*\text{A34}\end{cases}$$
$$+ \begin{cases}(-1.623\text{e}+00)*\text{A41}\\(-1.522\text{e}+00)*\text{A410}\\(-6.874\text{e}-01)*\text{A42}\\(-8.808\text{e}-01)*\text{A43}\\(-5.065\text{e}-01)*\text{A44}\\(-1.319\text{e}-01)*\text{A45}\\(1.272\text{e}-01)*\text{A46}\\(-2.124\text{e}+00)*\text{A48}\\(-8.061\text{e}-01)*\text{A49}\end{cases} + (1.114\text{e}-04)*amount + \begin{cases}(-2.960\text{e}-01)*\text{A62}\\(-4.453\text{e}-01)*\text{A63}\\(-1.341\text{e}+00)*\text{A64}\\(-9.750\text{e}-01)*\text{A65}\end{cases} + (3.098\text{e}-01)$$
$$*installment + \begin{cases}(-2.148\text{e}-01)*A92\\(-7.826\text{e}-01)*A93\\(-3.313\text{e}-01)*\text{A94}\end{cases} + \begin{cases}(4.934\text{e}-01)*A102\\(-1.036\text{e}+00)*A103\end{cases} + \begin{cases}(-6.513\text{e}-02)*A142\\(-6.709\text{e}-01)*A143\end{cases}$$
$$+ \begin{cases}(-4.864\text{e}-01)*A152\\(-3.338\text{e}-01)*A153\end{cases} + (-1.307\text{e}+00)*A202$$

Bellow shows the 95% confidence intervals of the coefficinets.

```
> confint(fit7)
                                  2.5 %        97.5 %
(Intercept)                -6.178950e-02  2.5482176059
factor(checkingstatus1)A12 -8.011958e-01  0.0267001583
factor(checkingstatus1)A13 -1.764327e+00 -0.3477317760
factor(checkingstatus1)A14 -2.189919e+00 -1.2975050695
duration                    1.102172e-02  0.0459203899
factor(history)A31         -1.165148e+00  0.8936704224
factor(history)A32         -1.658487e+00 -0.0499225687
factor(history)A33         -1.896883e+00 -0.0683817682
factor(history)A34         -2.407668e+00 -0.7146053298
factor(purpose)A41         -2.362974e+00 -0.9300589779
factor(purpose)A410        -3.089199e+00 -0.1056252269
factor(purpose)A42         -1.183395e+00 -0.1992467894
factor(purpose)A43         -1.357934e+00 -0.4105893961
factor(purpose)A44         -2.024318e+00  0.9089553459
factor(purpose)A45         -1.220855e+00  0.9098526045
factor(purpose)A46         -6.391272e-01  0.8897924739
factor(purpose)A48         -5.223160e+00 -0.0418720974
factor(purpose)A49         -1.456526e+00 -0.1744030082
amount                      3.143646e-05  0.0001927769
factor(savings)A62         -8.454965e-01  0.2372014365
factor(savings)A63         -1.257140e+00  0.2874299909
factor(savings)A64         -2.426408e+00 -0.4126719769
factor(savings)A65         -1.487594e+00 -0.4862281581
installment                 1.452499e-01  0.4780798823
factor(status)A92          -9.390854e-01  0.5179774157
factor(status)A93          -1.492417e+00 -0.0656751442
factor(status)A94          -1.190415e+00  0.5255775262
factor(others)A102         -3.021091e-01  1.2844939734
factor(others)A103         -1.897349e+00 -0.2540354521
factor(otherplans)A142     -8.603278e-01  0.7210284117
factor(otherplans)A143     -1.129921e+00 -0.2092746057
factor(housing)A152        -9.163301e-01 -0.0550656331
factor(housing)A153        -9.740098e-01  0.3018275739
factor(foreign)A202        -2.674464e+00 -0.1950849899
```

Problem 1 Section 2

```
### problem 1. a)

german_credit <- read.csv("germancredit.csv", header = T)

variables <- names(german_credit)
str(german_credit)

# 13 categorical and 7 numerical predictors for 1 response variable Default

attach(german_credit)

table(Default)

### problem 1. b)
chi_value <- matrix(nrow = 20, ncol = 1)
par(mfrow=c(2,10))
for (i in 2:21) {
  ta1 <- table(german_credit[,i], Default)
  barplot(prop.table(ta1, mar = 2), ylab = variables[i], xlab =variables[1])
  # chisq.test(ta1)
  fit1 <- glm(Default ~ german_credit[,i], family = binomial, data = german_credit) # fit glm for
each predictor
  # summary(fit1)
  anova_table <- anova(fit1, test = "Chisq")
  chi_value[i-1,1] <- anova_table$'Pr(>Chi)'[2] # Extracting chisqure value from anova_table
}


predictors.tables <- data.frame(variables[2:21], chi_value, ifelse(chi_value < 0.25, "ueseful",
"unuseful"))
colnames(predictors.tables) <- c("respond", "Chi-value", "decision")
predictors.tables
#            respond    Chi-value decision
# 1  checkingstatus1 2.787203e-28  ueseful
# 2         duration 2.398744e-11  ueseful
# 3          history 2.313958e-12  ueseful
# 4          purpose 7.268797e-05  ueseful
# 5           amount 1.928791e-06  ueseful
# 6          savings 7.049052e-08  ueseful
# 7           employ 1.146430e-03  ueseful
# 8      installment 2.116372e-02  ueseful
# 9           status 2.396272e-02  ueseful
# 10          others 3.597102e-02  ueseful
# 11       residence 9.252355e-01 unuseful
# 12        property 3.106302e-05  ueseful
# 13             age 3.389931e-03  ueseful
# 14      otherplans 2.129794e-03  ueseful
# 15         housing 1.448466e-04  ueseful
# 16           cards 1.443321e-01  ueseful
# 17             job 6.032614e-01 unuseful
# 18          liable 9.239826e-01 unuseful
# 19            tele 2.477554e-01  ueseful
# 20         foreign 4.494505e-03  ueseful

fit.full <- glm(Default ~ . , family = binomial, data = german_credit)
summary(fit.full)
```

```
fit2 <- glm(Default ~ factor(checkingstatus1) + duration + factor(history) + factor(purpose) +
amount +
              factor(savings) + factor(employ) + installment + factor(status) + factor(others) +
              factor(property) + age + factor(otherplans) + factor(housing) + cards +
              factor(tele) + factor(foreign), family = binomial, data = german_credit)
anova(fit.full, fit2, test = "Chisq")
summary(fit2)

# Remove employ
fit3 <- glm(Default ~ factor(checkingstatus1) + duration + factor(history) + factor(purpose) +
amount +
              factor(savings)  + installment + factor(status) + factor(others) +
              factor(property) + age + factor(otherplans) + factor(housing) + cards +
              factor(tele) + factor(foreign), family = binomial, data = german_credit)
summary(fit3)
anova(fit.full, fit3, test = "Chisq")

# Remove property
fit4 <- glm(Default ~ factor(checkingstatus1) + duration + factor(history) + factor(purpose) +
amount +
              factor(savings)  + installment + factor(status) + factor(others) +
              age + factor(otherplans) + factor(housing) + cards +
              factor(tele) + factor(foreign), family = binomial, data = german_credit)
summary(fit4)
anova(fit.full, fit4, test = "Chisq")

# Remove age
fit5 <- glm(Default ~ factor(checkingstatus1) + duration + factor(history) + factor(purpose) +
amount +
              factor(savings)  + installment + factor(status) + factor(others) +
              factor(otherplans) + factor(housing) + cards +
              factor(tele) + factor(foreign), family = binomial, data = german_credit)
summary(fit5)
anova(fit.full, fit5, test = "Chisq")

# Remove cards
fit6 <- glm(Default ~ factor(checkingstatus1) + duration + factor(history) + factor(purpose) +
amount +
              factor(savings) + installment + factor(status) + factor(others) +
              factor(otherplans) + factor(housing) +
              factor(tele) + factor(foreign), family = binomial, data = german_credit)
summary(fit6)
anova(fit.full, fit6, test = "Chisq")

# Remove tele
fit7 <- glm(Default ~ factor(checkingstatus1) + duration + factor(history) + factor(purpose) +
amount +
              factor(savings) + installment + factor(status) + factor(others) +
              factor(otherplans) + factor(housing) +
              factor(foreign), family = binomial, data = german_credit)
summary(fit7)
anova(fit.full, fit7, test = "Chisq")

### problem 1. c)
confint(fit7)
```