Exploratory Data Analysis

Zhiquan Cui

2023-03-25

Contents

1 Load Required Libraries

```
library(dplyr)
library(insight)
library(knitr)
library(kableExtra)
library(ggplot2)
library(tidyverse)
library(corrplot)
library(patchwork)
library(rcompanion)
```

2 Load Data & Inspect Variables

```
# Read the data
data <- read.csv("Credit.csv")
# Check the number of observations and number of variables
n <- nrow(data)
m <- ncol(data)
n
```

```
## [1] 1000
```

```
m
```

```
## [1] 21
```

```
# Check the data
kable(head(data[, 1:8]), format = "latex", align=rep("c", 8))
```

status	duration	credit_history	purpose	amount	savings	employment_duration	installment_rate
1	18	4	2	1049	1	2	4
1	9	4	0	2799	1	3	2
2	12	2	9	841	2	4	2
1	12	4	0	2122	1	3	3
1	12	4	0	2171	1	3	4
1	10	4	0	2241	1	2	1

```
kable(head(data[, 9:14]), format = "latex", align=rep("c", 6))
```

personal_status_sex	other_debtors	present_residence	property	age	other_installment_plans
2	1	4	2	21	3
3	1	2	1	36	3
2	1	4	1	23	3
3	1	2	1	39	3
3	1	4	2	38	1
3	1	3	1	48	3

```r
kable(head(data[, 15:21]), format = "latex", align=rep("c", 7))
```

housing	number_credits	job	people_liable	telephone	foreign_worker	credit_risk
1	1	3	2	1	2	1
1	2	3	1	1	2	1
1	1	2	2	1	2	1
1	2	2	1	1	1	1
2	2	2	2	1	1	1
1	2	2	1	1	1	1

```r
# Check invalid or missing values
anyNA(data)
```

```
## [1] FALSE
```

```r
# Check the data type of each column
sapply(data, class)
```

```
##                status               duration         credit_history
##             "integer"              "integer"              "integer"
##               purpose                 amount                savings
##             "integer"              "integer"              "integer"
##   employment_duration       installment_rate    personal_status_sex
##             "integer"              "integer"              "integer"
##         other_debtors      present_residence               property
##             "integer"              "integer"              "integer"
##                   age other_installment_plans               housing
##             "integer"              "integer"              "integer"
##        number_credits                    job          people_liable
##             "integer"              "integer"              "integer"
##             telephone         foreign_worker            credit_risk
##             "integer"              "integer"              "integer"
```

As we can see from the above outputs, there is no NaN values so the data is clean. And all of the columns are of type integer. Some of them are quantitative variable while some of them are qualitative variables. Here is a summary of the variables:

- status: status of the debtor's checking account with the bank (categorical)
- duration: credit duration in months (quantitative)
- credit_history: history of compliance with previous or concurrent credit contracts (categorical)
- purpose: purpose for which the credit is needed (categorical)
- amount: credit amount in DM (quantitative; result of monotonic transformation; actual data and type of transformation unknown)
- savings: debtor's savings (categorical)
- employment_duration: duration of debtor's employment with current employer (ordinal; discretized quantitative)
- installment_rate: credit installments as a percentage of debtor's disposable income (ordinal; discretized quantitative)
- personal_status_sex: combined information on sex and marital status (categorical)
- other_debtors: is there another debtor or a guarantor for the credit? (categorial)
- present_residence: length of time (in years) the debtor lives in the present residence (ordinal; discretized quantitative)

- property: the debtor's most valuable property (ordinal)
- age: age in years (quantitative)
- other_installment_plans: installment plans from providers other than the credit-giving bank (categorical)
- housing: type of housing the debtor lives in (categorical)
- number_credits: number of credits including the current one the debtor has (or had) at the bank (ordinal; discretized quantitative)
- job: quality of debtor's job (ordinal)
- people_liable: number of persons who financially depend on the debtor (binary; discretized quantitative)
- telephone: is there a telephone landline registered on the debtor's name? (binary)
- foreign_ worker: is the debtor a foreign worker? (binary)
- credit_risk: has the credit contract been complied with (good) or not (bad)? (binary)

We can see that the **quantitative variables** include duration, amount and age, while **qualitative variables** include status, credit_history, purpose, savings, employment_duration, installment_rate, personal_status_sex, other_debtors, present_residence, property, other_installment_plans, housing, number_credits, job, people_liable, telephone, foreign_worker and credit_risk.

3 Univariate Data Analysis & Visualization

3.1 Histogram of Quantitative Variables

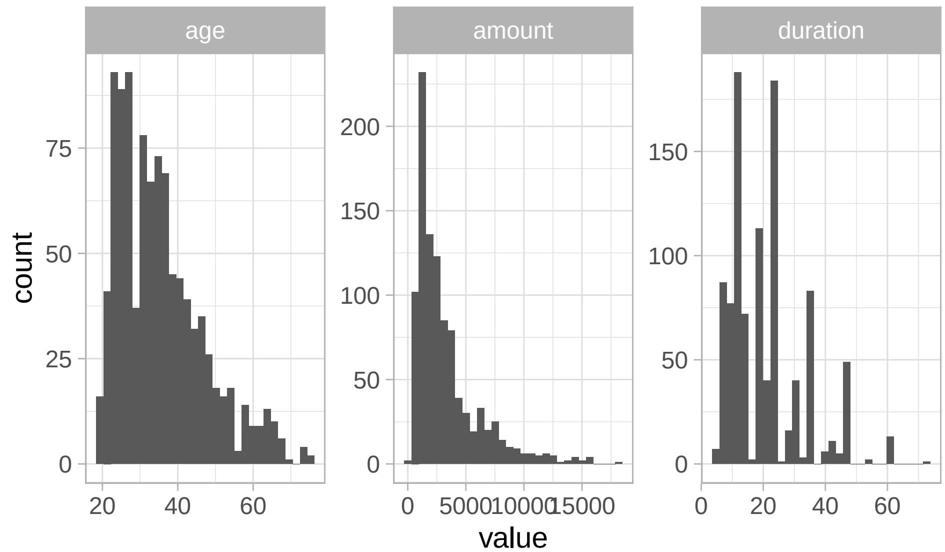
First we will perform univariate analysis on each of the variables and look at their distribution. Here is the summary statistics:

```
quant_vars <- c("duration", "amount", "age")
qual_vars <- c("status", "credit_history", "purpose", "savings", "employment_duration",
               "installment_rate", "personal_status_sex", "other_debtors", "present_residence",
               "property", "other_installment_plans", "housing", "number_credits", "job",
               "people_liable", "telephone", "foreign_worker")
summary(data[, quant_vars])
```

```
##     duration        amount           age
##  Min.   : 4.0   Min.   :  250   Min.   :19.00
##  1st Qu.:12.0   1st Qu.: 1366   1st Qu.:27.00
##  Median :18.0   Median : 2320   Median :33.00
##  Mean   :20.9   Mean   : 3271   Mean   :35.54
##  3rd Qu.:24.0   3rd Qu.: 3972   3rd Qu.:42.00
##  Max.   :72.0   Max.   :18424   Max.   :75.00
```

Next, let us check the histograms of the quantitative variables:

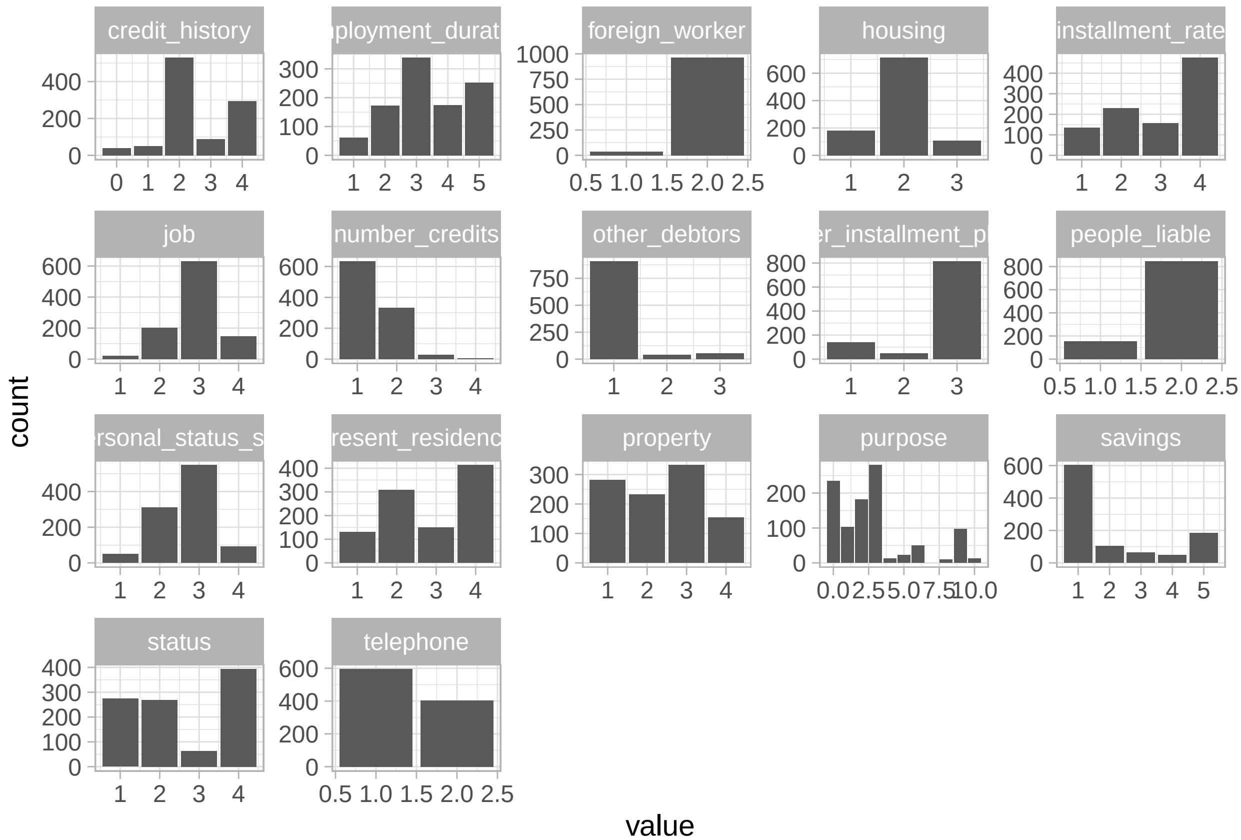
```
data[, quant_vars] %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram() +
    theme_light()
```



3.2 Barplot of Qualitative Variables

Then, let us check the barplots of qualitative variables:

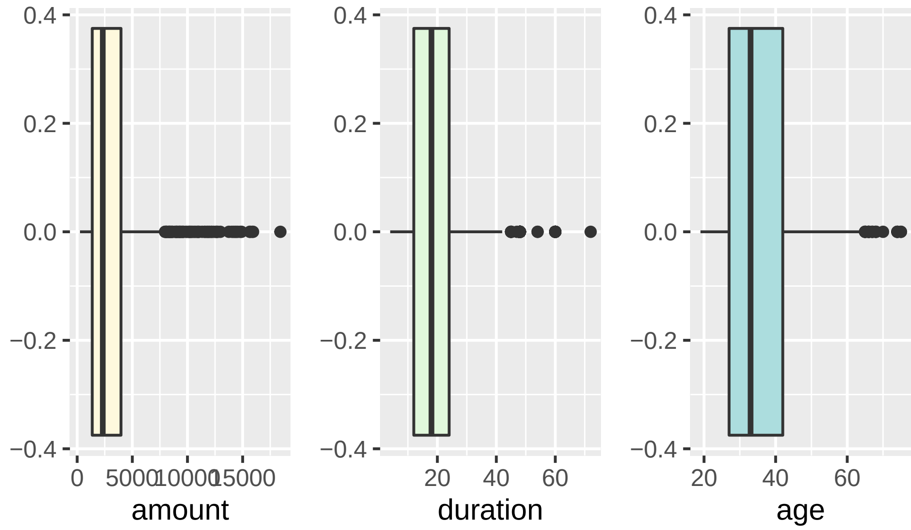
```
data[, qual_vars] %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_bar() +
    theme_light()
```



3.3 Boxplot of Quantitative Variables

After checking the histograms and barplots, we will check the boxplots of the quantitative variables. Here we will not check barplots for qualitative variables because it only makes sense to examine the median, first and third quartiles and maximum value for quantitative variables.

```
g1 <- ggplot(data, aes(x = amount)) + geom_boxplot(fill="#FEF8DD")
g2 <- ggplot(data, aes(x = duration)) + geom_boxplot(fill="#E1F8DC")
g3 <- ggplot(data, aes(x = age)) + geom_boxplot(fill="#ACDDDE")
g1 + g2 + g3
```



From the above box plots, we can see that there are a few outliers for the variable amount. If we look at the histogram of variable amount, we can see that it is a right skewed distribution with a long right tail, which results in these outliers.

3.4 Sample Odds of Binary Variables

For binary variables people_liable, telephone, foreign_worker and credit_risk, we can calculate and interpret the sample odds:

```
binary_var <- c("Statistics", "people_liable", "telephone", "foreign_worker", "credit_risk")
odds <- c("Sample Odds")
for (var in binary_var[2:5]) {
  if (var == "credit_risk") {y <- sum(data[, var] == 1)}
  else {y <- sum(data[, var] == 2)}
  n <- length(data[, var])
  odds <- append(odds, round(y / (n - y), 2))
}
kable(data.frame(t(odds)), col.names = binary_var, format = "latex") %>%
  kable_styling(position = "center", latex_options = "hold_position") %>% row_spec(0, bold = TRUE)
```

Statistics	**people_liable**	**telephone**	**foreign_worker**	**credit_risk**
Sample Odds	5.45	0.68	26.03	2.33

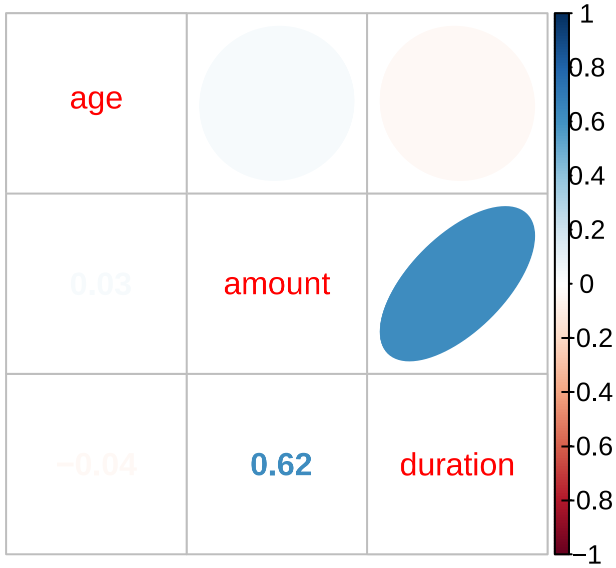
Based on our sample, the estimated probability of a person to have good credit is 2.33 times as likely as having a bad credit. Similarly, the estimated probability of a person to have a telephone landline registered on his/her name is 0.68 times as likely as not having such a telephone landline.

4 Multivariate Data Analysis & Visualization

4.1 Quantitative Variable

First, let us look at the correlation plots of the quantitative variables.

```r
corrplot.mixed(cor(data[quant_vars]), lower='number', upper='ellipse', order='AOE')
```



From the above correlation plot, we can see that the correlation coefficient between amount and duration is as high as 0.62, which indicates a strong positive correlation between the two variables. This also makes sense intuitively because the longer credit duration one has in months, he/she will have a higher chance to build up his/her credit and obtain a higher credit amount. Similarly, if one has a high credit amount, then he/she is more likely to have a long credit duration. In order to avoid multicollinearity, we will drop the variable amount in our model.

4.2 Qualitative Variables

After examining the quantitative variables, we will now look at the qualitative variables. Since they are not continuous and numeric data, we should not use the same methodology as above. Instead, we will use Pearson's Chi-sq Test of Indepence and Cramer's V designed for qualitative variables to examine the data.

```r
Pearson_chisq_test <- data.frame(matrix(0, ncol = length(qual_vars),
                                nrow = length(qual_vars)), row.names = qual_vars)
colnames(Pearson_chisq_test) <- qual_vars
for (var in qual_vars) {
  for (var_2 in qual_vars) {
    test <- chisq.test(table(data[, var], data[, var_2]), simulate.p.value = TRUE)
    Pearson_chisq_test[var, var_2] <- test$p.value
  }
}
kable(Pearson_chisq_test[, 1:5], format = "latex")
```

	status	credit_history	purpose	savings	employment_duration
status	0.0004998	0.0004998	0.0004998	0.0004998	0.0079960
credit_history	0.0004998	0.0004998	0.0004998	0.2208896	0.0019990
purpose	0.0004998	0.0004998	0.0004998	0.0499750	0.0149925
savings	0.0004998	0.1959020	0.0509745	0.0004998	0.0309845
employment_duration	0.0059970	0.0004998	0.0119940	0.0254873	0.0004998
installment_rate	0.4372814	0.6926537	0.0009995	0.3508246	0.0004998
personal_status_sex	0.1414293	0.0159920	0.0004998	0.4642679	0.0004998
other_debtors	0.0014993	0.0514743	0.0009995	0.0194903	0.0879560
present_residence	0.0019990	0.1094453	0.0044978	0.1424288	0.0004998
property	0.0454773	0.0929535	0.0004998	0.0764618	0.0004998
other_installment_plans	0.2688656	0.0004998	0.0034983	0.9995002	0.2698651
housing	0.0029985	0.0174913	0.0004998	0.8260870	0.0004998
number_credits	0.0374813	0.0004998	0.2693653	0.1074463	0.0004998
job	0.0474763	0.3903048	0.0004998	0.3433283	0.0004998
people_liable	0.1144428	0.0554723	0.0049975	0.8850575	0.0544728
telephone	0.0879560	0.2778611	0.0004998	0.0619690	0.0004998
foreign_worker	0.1264368	0.4037981	0.0059970	0.9045477	0.4792604

```
kable(Pearson_chisq_test[, 6:10], format = "latex")
```

	installment_rate	personal_status_sex	other_debtors	present_residence	property
status	0.4472764	0.1569215	0.0019990	0.0009995	0.0469765
credit_history	0.6811594	0.0109945	0.0589705	0.1074463	0.0799600
purpose	0.0014993	0.0014993	0.0019990	0.0059970	0.0004998
savings	0.3338331	0.4647676	0.0164918	0.1364318	0.0924538
employment_duration	0.0004998	0.0004998	0.0779610	0.0004998	0.0004998
installment_rate	0.0004998	0.0004998	0.9790105	0.4212894	0.4427786
personal_status_sex	0.0004998	0.0004998	0.6051974	0.0009995	0.0004998
other_debtors	0.9770115	0.5967016	0.0004998	0.6371814	0.0004998
present_residence	0.4122939	0.0009995	0.6416792	0.0004998	0.0004998
property	0.4532734	0.0004998	0.0004998	0.0004998	0.0004998
other_installment_plans	0.6186907	0.4387806	0.2248876	0.7366317	0.0039980
housing	0.1079460	0.0004998	0.1289355	0.0004998	0.0004998
number_credits	0.4142929	0.0559720	0.8335832	0.0114943	0.1149425
job	0.0649675	0.0389805	0.0494753	0.9050475	0.0004998
people_liable	0.1209395	0.0004998	0.3533233	0.2773613	0.0374813
telephone	0.4747626	0.0464768	0.0594703	0.0164918	0.0004998
foreign_worker	0.0024988	0.0809595	0.0019990	0.3783108	0.0004998

```
kable(Pearson_chisq_test[, 11:15], format = "latex")
```

	other_installment_plans	housing	number_credits	job	people_liable
status	0.2923538	0.0029985	0.0339830	0.0479760	0.1104448
credit_history	0.0004998	0.0179910	0.0004998	0.3823088	0.0504748
purpose	0.0029985	0.0004998	0.2978511	0.0004998	0.0014993
savings	0.9985007	0.8245877	0.1019490	0.3438281	0.8970515
employment_duration	0.2848576	0.0004998	0.0014993	0.0004998	0.0499750
installment_rate	0.6251874	0.1274363	0.4162919	0.0574713	0.1034483
personal_status_sex	0.4542729	0.0004998	0.0509745	0.0429785	0.0004998
other_debtors	0.2253873	0.1289355	0.8485757	0.0544728	0.3358321
present_residence	0.7296352	0.0004998	0.0104948	0.8975512	0.2738631
property	0.0009995	0.0004998	0.1214393	0.0004998	0.0334833
other_installment_plans	0.0004998	0.0034983	0.0224888	0.0329835	0.0479760
housing	0.0059970	0.0004998	0.0049975	0.0004998	0.0014993
number_credits	0.0184908	0.0064968	0.0004998	0.0024988	0.0059970
job	0.0304848	0.0004998	0.0029985	0.0004998	0.0004998
people_liable	0.0449775	0.0004998	0.0064968	0.0004998	0.0004998
telephone	0.2678661	0.0004998	0.0449775	0.0004998	0.6436782
foreign_worker	0.8365817	0.0309845	0.9210395	0.0184908	0.0199900

```r
kable(Pearson_chisq_test[, 16:17], format = "latex")
```

	telephone	foreign_worker
status	0.1064468	0.1274363
credit_history	0.2658671	0.4137931
purpose	0.0004998	0.0079960
savings	0.0699650	0.9145427
employment_duration	0.0004998	0.4602699
installment_rate	0.4712644	0.0059970
personal_status_sex	0.0469765	0.0779610
other_debtors	0.0479760	0.0019990
present_residence	0.0269865	0.3673163
property	0.0004998	0.0004998
other_installment_plans	0.2728636	0.8430785
housing	0.0029985	0.0274863
number_credits	0.0599700	0.9110445
job	0.0004998	0.0199900
people_liable	0.6511744	0.0219890
telephone	0.0004998	0.0259870
foreign_worker	0.0309845	0.0004998

Based on the above table, we conclude that the following variables are dependent to most of the variables with $\alpha = 0.05$ according to Pearson's Chi-sq Test of Independence:

- purpose
- employment_duration
- property
- housing
- job
- people_liable
- number_credits
- credit_history

```r
Cramer_v <- data.frame(matrix(0, ncol = length(qual_vars),
                       nrow = length(qual_vars)), row.names = qual_vars)
colnames(Cramer_v) <- qual_vars
for (var in qual_vars) {
  for (var_2 in qual_vars) {
    Cramer_v[var, var_2] <- cramerV(table(data[, var], data[, var_2]))
  }
}
kable(Cramer_v[, 1:5], format = "latex")
```

	status	credit_history	purpose	savings	employment_duration
status	1.00000	0.14180	0.1492	0.17560	0.09532
credit_history	0.14180	1.00000	0.1671	0.07151	0.10070
purpose	0.14920	0.16710	1.0000	0.11380	0.12170
savings	0.17560	0.07151	0.1138	1.00000	0.08465
employment_duration	0.09532	0.10070	0.1217	0.08465	1.00000
installment_rate	0.05457	0.05526	0.1396	0.06689	0.10840
personal_status_sex	0.06717	0.09242	0.1511	0.06226	0.16620
other_debtors	0.10850	0.08751	0.1653	0.09651	0.08318
present_residence	0.09811	0.07757	0.1299	0.07615	0.26140
property	0.07576	0.07982	0.2058	0.07911	0.14710
other_installment_plans	0.06111	0.26550	0.1443	0.02021	0.06960
housing	0.09887	0.09650	0.2067	0.04655	0.16980
number_credits	0.07695	0.37820	0.1014	0.07851	0.11580
job	0.07518	0.06485	0.2028	0.06646	0.31130
people_liable	0.07694	0.09769	0.1637	0.03391	0.09799
telephone	0.08091	0.07152	0.2206	0.09306	0.15060
foreign_worker	0.07593	0.06313	0.1711	0.03263	0.05963

```r
kable(Cramer_v[, 6:10], format = "latex")
```

	installment_rate	personal_status_sex	other_debtors	present_residence	property
status	0.05457	0.06717	0.10850	0.09811	0.07576
credit_history	0.05526	0.09242	0.08751	0.07757	0.07982
purpose	0.13960	0.15110	0.16530	0.12990	0.20580
savings	0.06689	0.06226	0.09651	0.07615	0.07911
employment_duration	0.10840	0.16620	0.08318	0.26140	0.14710
installment_rate	1.00000	0.10250	0.02473	0.05518	0.05452
personal_status_sex	0.10250	1.00000	0.04744	0.10580	0.12010
other_debtors	0.02473	0.04744	1.00000	0.04624	0.14210
present_residence	0.05518	0.10580	0.04624	1.00000	0.13600
property	0.05452	0.12010	0.14210	0.13600	1.00000
other_installment_plans	0.04698	0.05369	0.05320	0.04231	0.09798
housing	0.07127	0.20150	0.05886	0.23560	0.55000
number_credits	0.05504	0.07614	0.03663	0.08644	0.06773
job	0.07328	0.07714	0.08045	0.03728	0.19390
people_liable	0.07823	0.28430	0.04801	0.06210	0.09477
telephone	0.05062	0.08916	0.07609	0.09806	0.19780
foreign_worker	0.11650	0.08155	0.14100	0.05570	0.14930

```r
kable(Cramer_v[, 11:15], format = "latex")
```

	other_installment_plans	housing	number_credits	job	people_liable
status	0.06111	0.09887	0.07695	0.07518	0.07694
credit_history	0.26550	0.09650	0.37820	0.06485	0.09769
purpose	0.14430	0.20670	0.10140	0.20280	0.16370
savings	0.02021	0.04655	0.07851	0.06646	0.03391
employment_duration	0.06960	0.16980	0.11580	0.31130	0.09799
installment_rate	0.04698	0.07127	0.05504	0.07328	0.07823
personal_status_sex	0.05369	0.20150	0.07614	0.07714	0.28430
other_debtors	0.05320	0.05886	0.03663	0.08045	0.04801
present_residence	0.04231	0.23560	0.08644	0.03728	0.06210
property	0.09798	0.55000	0.06773	0.19390	0.09477
other_installment_plans	1.00000	0.09285	0.09204	0.08401	0.07722
housing	0.09285	1.00000	0.09962	0.12320	0.12780
number_credits	0.09204	0.09962	1.00000	0.11160	0.12070
job	0.08401	0.12320	0.11160	1.00000	0.14600
people_liable	0.07722	0.12780	0.12070	0.14600	1.00000
telephone	0.05061	0.11510	0.08494	0.42570	0.01475
foreign_worker	0.01890	0.08439	0.02212	0.10160	0.07707

```
kable(Cramer_v[, 16:17], format = "latex")
```

	telephone	foreign_worker
status	0.08091	0.07593
credit_history	0.07152	0.06313
purpose	0.22060	0.17110
savings	0.09306	0.03263
employment_duration	0.15060	0.05963
installment_rate	0.05062	0.11650
personal_status_sex	0.08916	0.08155
other_debtors	0.07609	0.14100
present_residence	0.09806	0.05570
property	0.19780	0.14930
other_installment_plans	0.05061	0.01890
housing	0.11510	0.08439
number_credits	0.08494	0.02212
job	0.42570	0.10160
people_liable	0.01475	0.07707
telephone	1.00000	0.07501
foreign_worker	0.07501	1.00000

Based on the above Cramer's V table, we conclude that we will drop the following variables because of their high correlation to other variables:

- job
- credit_history
- purpose
- employment_duration
- housing
- people_liable

To summarize, the variables we will use in model building are:

- status

- duration
- savings
- employment_duration
- installment_rate
- personal_status_sex
- other_debtors
- present_residence
- property
- age
- other_installment_plans
- number_credit
- telephone
- foreign_worker