Exploratory Analysis of the South German Credit Dataset (Updated in 2019 by U. Grömping)

Gregory Pollard

March 21, 2021

1 Describing the Problem

Thoughout this report (and its successor), we'll attempt to answer the following research question:

"Provided with the 2019 revised version of the "South German Credit" (SGC) dataset, can we create a suitable statistical model to predict if a new customer's credit risk will be good or bad?"

In this context (and for the remainder of this report), 'good' implies that the customer will/has comply/complied with all conditions of the contract, and 'bad' means that they may/have breach/breached the terms of their contract. This report will focus heavily on the data understanding required for this problem. A thorough investigation into the data itself has already been conducted and will be referred to throughout this report also, it consists of a technical report published in April 2019 titled "South German Credit Data: Correcting a Widely Used Data Set" [1]. The eventual goal is to produce a classification model capable of deciding whether a new customer is credit-worthy.

2 Data Structure and Understanding

Data on its own is useless for prediction. On their own, tables with numbers and strings in each cell are not helpful if we don't know what these values represent. Fortunately, in the SGC data we have 20 columns with variable names attached to each column labelling the relevant data on a particular aspect of each customer in the database (e.g. age). A table of all variables and what they represent in context can be found in the cited report [2]. There are exactly 1,000 records in the database, 18 categorical variables and only 3 quantitative. A list of all variable names and their corresponding ranges/levels in context is stated in [1].

There are exactly 300 records with a 'bad' credit risk, and exactly 700 records with a 'good' credit risk. From [1] we know that this was intentional since only about 5% of customers are ever categorised as 'bad'. This is exactly why we **must not** compare the **frequencies** of our explanatory variables with the credit risk variable since the proportion of customers in the dataset with a 'bad' credit risk is **not** representative of the population. Our study focuses on why one might be classified as 'bad', and so to make a comparison to those who are 'good', we need to have many records of 'bad' customers (which is why there are many more 'bad' customers than 'good' in this dataset).

3 Impressions on Data Quality Prior to Analysis and Model Choice

Before we explore any relationships in the data we must check for missing values, are there any cells that have no valid input? Or perhaps have no value at all? Reference [2] details some metadata regarding our dataset that clearly states no N/A values exist. This dataset was examined and formatted by U.Grömping and so is unlikely to contain any impactful errors (if any). However, outliers may still be present. In addition to this, we need to decide which variables may impact credit risk enough to be considered later on in model fitting, it could be the case that the information provided by a particular variable is accounted for by another and thus we may not need to include both.

We need to explore some ideas on how these variables may relate to each other and credit risk, this way we can attempt to cut out variables that aren't very impactful on credit risk before crafting a proper model. We can use some simple visualisation techniques such as relationship diagrams, scatter plots, and bar graphs to explore relationships in the data. It is important to do this so that:

- We do not have a model that is too complex.
- We only choose variables that affect the response variable considerably.
- We can create an overview of the data's basic properties.

Since we aim to predict a binary response variable, and we also have many explanatory variables of differing types and possibly different relationships to the response variable (e.g. an x^2 or $\log(x)$ relationships), we need to be careful with our model choice and exploratory analysis techniques. A good candidate would be to construct a generalised linear model (GLM) utilising a logistic regression approach. Using this method, we can incorporate many variables with their differing impacts and relationships to each other in one multivariate equation. In order to employ a GLM however, we need to make the following assumptions about our data:

1. The responses $(Y_i, i = 1, 2, \ldots, n)$ are realisations of random variables which are observed **independently**.
2. The conditional distribution Y is a member of the **exponential family**, where the conditioning is on the observed values of the explanatory variables.
3. The explanatory variables influence the distribution Y through a single linear function called the **linear predictor**:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

So, any given covariate x_j $(\forall j \in \{1, 2...p\})$ has no influence on the distribution of Y if and only if $\beta_j = 0$.

These assumptions seem rather fair to make, there's no obvious reason why the records wouldn't be independent. We can also temporarily assume a conditional distribution for now, and thus through our analysis determine if it actually exists.

Another model fitting technique we could utilise are decision trees, if our later analysis reveals that such a model would perform better than a GLM, then it may be more optimal. To forward this idea, decision trees aren't concerned with irrelevant/correlated variables and we don't need to worry about missing values or outliers, these points could in fact make a decision tree more attractive than a GLM.

4 Initial Thoughts and Significance of Variables

4.1 Visualisations of Preliminary Impressions on Relationships

In this section, we'll try to identify some of these relationships to help us decide which variables **could be** excluded from our model. Figure 1 is a relationship diagram based on our initial impressions of these variables and how one might assign relationships through intuition. In addition, it shows which variables are likely to increase and decrease credit risk, as well as showing how direct or indirect these relationships are via the number of arrows required to reach the *credit_risk* node. For example, one could expect the *credit_risk* to have a higher probability of being 'good' if a given debtor has an excellent history of compliance with the bank.

A red node represents a backward, one-way relationship **away** from the credit risk node (for example, having a landline doesn't **effect** the *housing* variable, but the reverse effect maybe true). This implies that the variable in question may not have a statistically significant impact on credit risk when considered in a finalised model. An orange node indicates a variable that is suspected to be independent of all other variables in the dataset, and thus it shouldn't be included. However, we must **prove** these claims in order to exclude them for our later analysis. Seeing as these are our initial impressions of the relationships, we should perform hypothesis tests for significance on *credit_risk* with **all** variables, and not simply those that we expect to be independent. The following subsection displays the results of these χ^2 hypothesis tests at the 95% significance.

NOTE: The *personal_status_sex* variable was not considered in this diagram and will not be used in analysis since (as stated in [2]): "sex cannot be recovered from the variable, because male singles and female non-singles are coded with the same code (2); female widows cannot be easily classified, because the code table does not list them in any of the female categories".

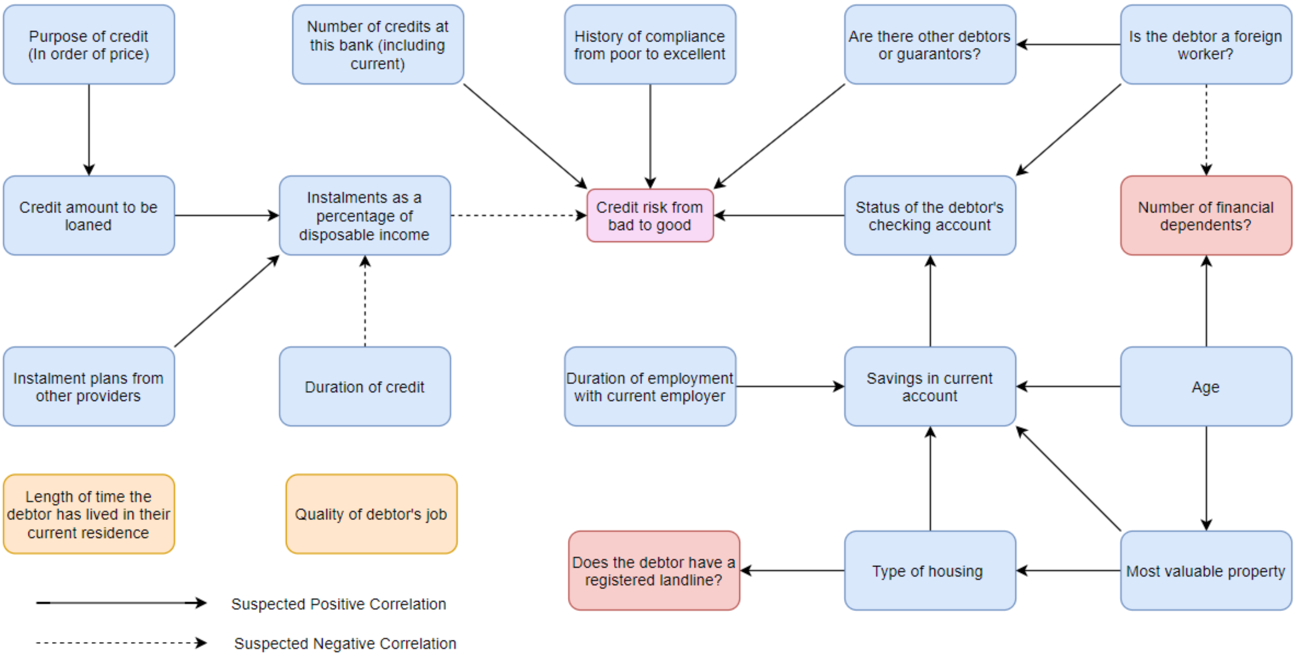
Purpose of credit (In order of price)
Number of credits at this bank (including current)
History of compliance from poor to excellent
Are there other debtors or guarantors?
Is the debtor a foreign worker?
Credit amount to be loaned
Instalments as a percentage of disposable income
Credit risk from bad to good
Status of the debtor's checking account
Number of financial dependents?
Instalment plans from other providers
Duration of credit
Duration of employment with current employer
Savings in current account
Age
Length of time the debtor has lived in their current residence
Quality of debtor's job
Does the debtor have a registered landline?
Type of housing
Most valuable property
Suspected Positive Correlation
Suspected Negative Correlation

Figure 1: A relationship diagram of all variables in our dataset.

4.2 Hypothesis Test Results

The following hypotheses were applied to all 20 variables:

H_0: *The [variable name] has no effect on credit risk.*
H_1: *The [variable name] has a statistically significant effect on credit risk.*

Variable Name	P-Value	Reject H_0?	Variable Name	P-Value	Reject H_0?
status	$1.218902e-26$	TRUE	other_installment_plans	0.001629318	TRUE
credit_history	$1.279187e-12$	TRUE	housing	$8.810311e-05$	TRUE
purpose	0.0001157491	TRUE	number_credits	0.4451441	FALSE
savings	$2.761214e-07$	TRUE	job	0.5965816	FALSE
employment_duration	0.001045452	TRUE	people_liable	1	FALSE
installment_rate	0.1400333	FALSE	telephone	0.01583075	TRUE
personal_status_sex	N/A	N/A	foreign_worker	0.01583075	TRUE
other_debtors	0.03605595	TRUE	duration	$1.218902e-26$	TRUE
present_residence	0.8615521	FALSE	amount	$1.218902e-26$	TRUE
property	$2.858442e-05$	TRUE	age	$1.218902e-26$	TRUE

Figure 2: A table of all variables, their p-values, and whether they have a significant effect on the *credit_risk* variable.

At this point it may be important to clarify what these results actually mean, to do this, we can interpret our p-values in the following way:

"Supposing the null hypothesis ***was*** *true, the probability of sampling data more extreme than our observed dataset is approximately [insert p-value here] with 95% confidence."*

From Figure 2 we can see that most of our variables seem to have a significant impact in some way or another to the credit risk; that is, their p-values are less than 0.05 (the standard for the degree of significance). However, we can also see that the *installment_rate, present_residence, number_credits, job,* and *people_liable* variables deviate from this trend.

4.3 Hypothesis Test Discussion

Although we concluded that some of our variables do **not** have a significant relationship with the *credit_risk* variable, it is a folly to assume that they are not of use at all. It may in fact be possible that a variable that appears to be uninformative is actually incredibly useful when combined with others in a fitted model; in contrast, an extremely low p-value may not necessarily indicate usefulness if there are other attributes in the same dataset that correlate closely with it. In this case, both of the variables would be just as informative as each other and thus we could hypothetically omit one from the analysis. Nonetheless, the *people_liable* attribute received a 1 for it's p-value and this **cannot** be ignored as it implies a total independence with respect to the *credit_risk* variable. This (in theory) means that; given we know the *people_liable* values, we might as well flip a coin to decide credit risk based on this information, thus we can safely omit this variable from the study.

5 Analysis of Relationships Between Variables

5.1 Relationships Between Numerical Variables

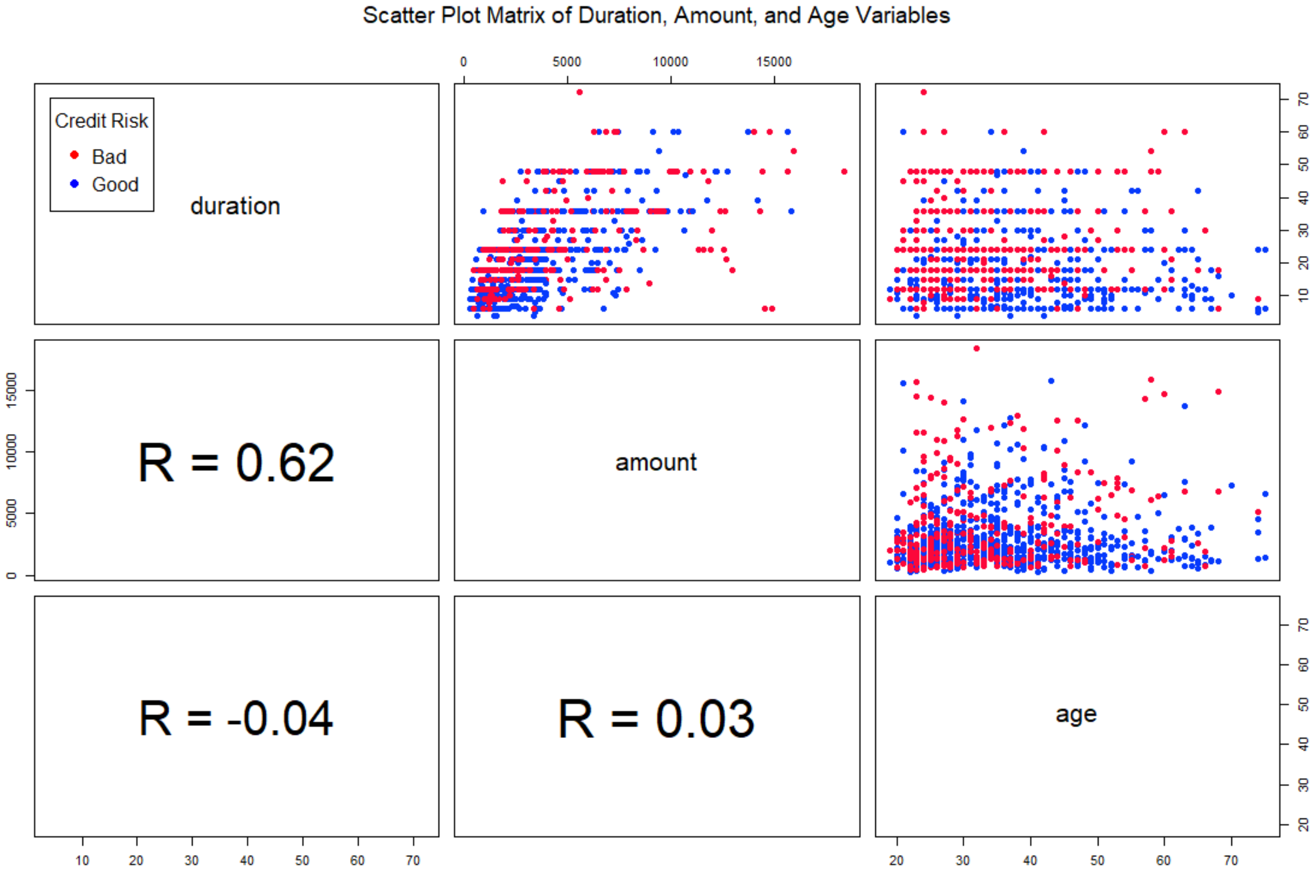

Figure 3: The letter R refers to the Pearson correlation coefficient. Note that we can see the 'bad' and 'good' points overlap considerably, implying that a decision tree **may not** work effectively as a classification model.

Figure 3 shows a scatter plot matrix of our 3 quantitative variables (*duration*, *amount*, and *age*). There are many relationships to examine here. We can see from the Pearson correlation coefficients in the lower panels that *age* has **no linear/polynomial correlation** with *duration* or *amount*. However, we can see from the plot that the majority of debtor's are young to middle-aged and generally have less than around 5000 DM in their account. We also have a moderate correlation between *duration* and *amount*; a relationship that can easily be seen in the plot itself. We can safely say that these relationships are not strong enough to omit any of the quantitative variables from our study, however we may find during model fitting that these values have varying levels of significance when seen as part of an actual GLM.

5.2 Relationships Between Categorical Variables

Before we detail these relationships, it's important to note that there were seemingly erroneous data discovered throughout the analysis of these categorical attributes, we will explain this in more detail in section 6. In this section, we will describe and visualise some significant relationships between our remaining 15 categorical explanatory variables. Consequently, we'll discuss what the relationships mean and their possible effects on model fitting later down the line.

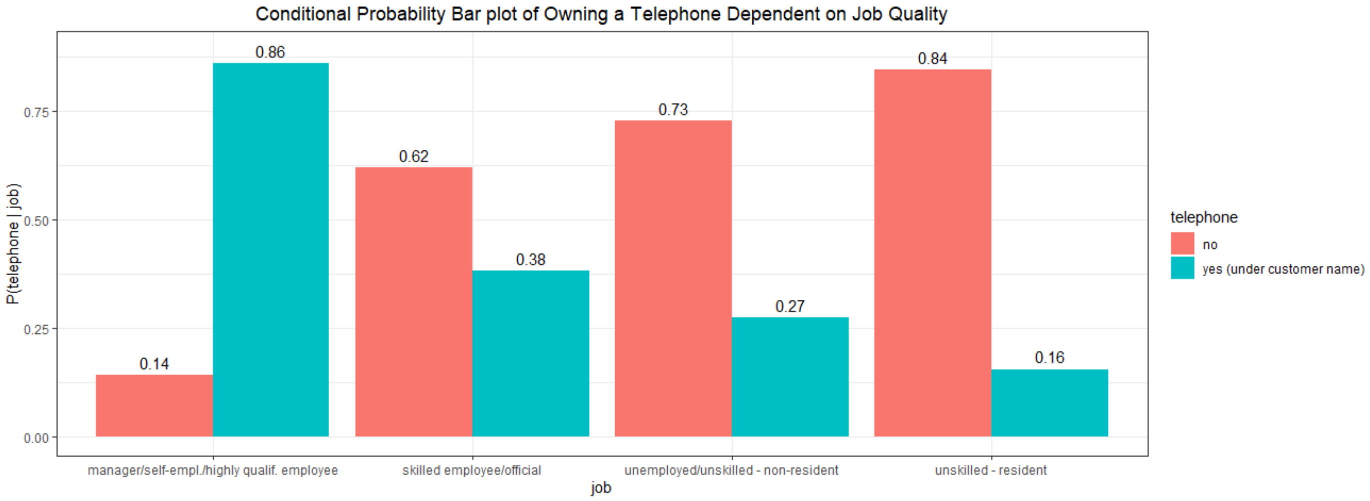

Figure 4: This conditional bar plot of *telephone* dependent on *job* has a clear significant relationship, however when conditioning in the opposing direction, this was not the case (implying a one-way relationship).

Figure 4 shows a conditional probability bar plot of the binary variable *telephone* on the *job* factor. The levels of *job* have been arranged in such a way that we can see the change in probability from an unemployed worker to those with managerial positions, this allows us to see how the ordinal nature of *job* effects whether or not there is a telephone landline registered in the debtor's name. As we can see, the probability of owning a telephone **increases** as the debtor's job quality also **increases**, this is a rather strong relationship and isn't that surprising when you consider that a highly skilled worker would most likely have a necessity to communicate by phone more often and thus require a landline. If we decide to include these in a model, we should look to add an interaction term to account for this (that is, if we can prove that adding this term would significantly improve our accuracy). The *number_credits* also has a similar relationship to *telephone* as the one described in Figure 4. Once again, we should consider an interaction term when model fitting.

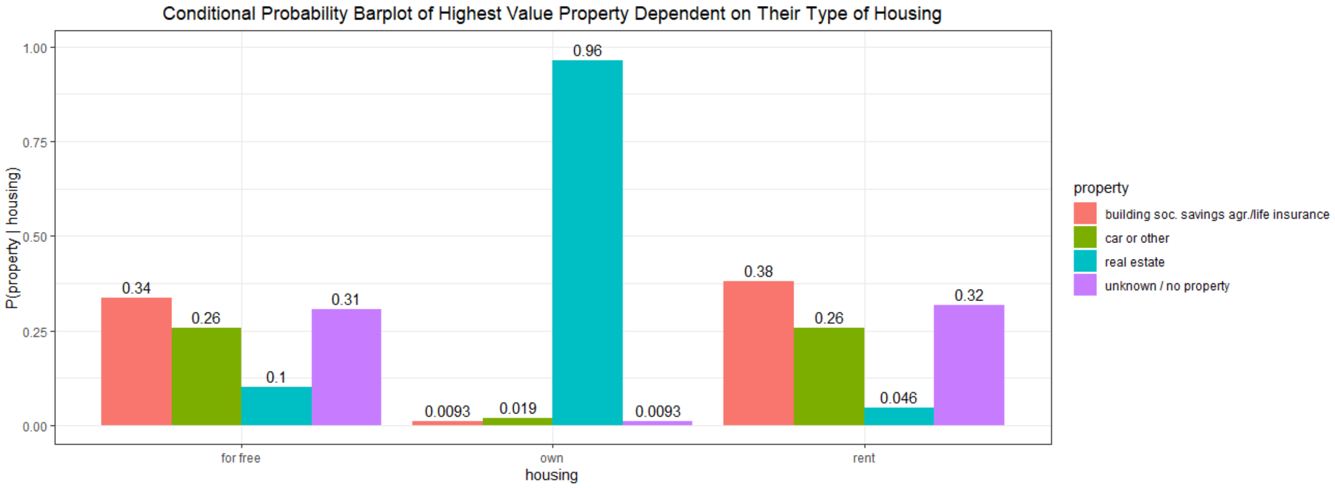

Figure 5: Unlike in Figure 4, conditioning in the opposite direction retains the same relationship, but to a lesser extent. Implying a weak relationship in both directions.

Figure 5 shows another conditional probability bar plot, this time with *property* dependent on *housing*. We can observe two major features of this graph at a glance, debtors who own their own residences are almost guaranteed to have this very same residence as the highest valued property they own (as shown by the tall blue bar). The second observation is that for the other two levels of *housing* ('for free' and 'rent'), we have fairly equal proportions across all levels of the *property* variable apart from 'real estate'. This implies that the debtor's residence is independent from what their highest valued property is, **if** they do **not** own their own home. This seems to make sense if we consider that those who do not pay or instead choose to pay rent are very likely to have less income as those who own their own home and pay a mortgage. These debtor's highest valued property would therefore be anything **but** real estate.

We should also be weary that there is an ‘unknown/no property’ level in the *property* factor. The ‘unknown’ and ‘no property’ elements being classified together is an issue since these unknown values could take up any proportion of this group, perhaps every record in this group is ‘unknown’? There is no way to tell, and so when fitting a model we should be very careful when including this attribute. In fact, discrepancies in the data like this cannot go unspoken; we’ll go into more detail about issues like this later on in section 6.

5.3 Relationships Between Categorical and Numerical Variables

Relative frequency conditional histograms allow us to shed light on any trends between our categorical and numerical data. Figure 6 displays the two strongest relationships uncovered utilising this visualisation method.

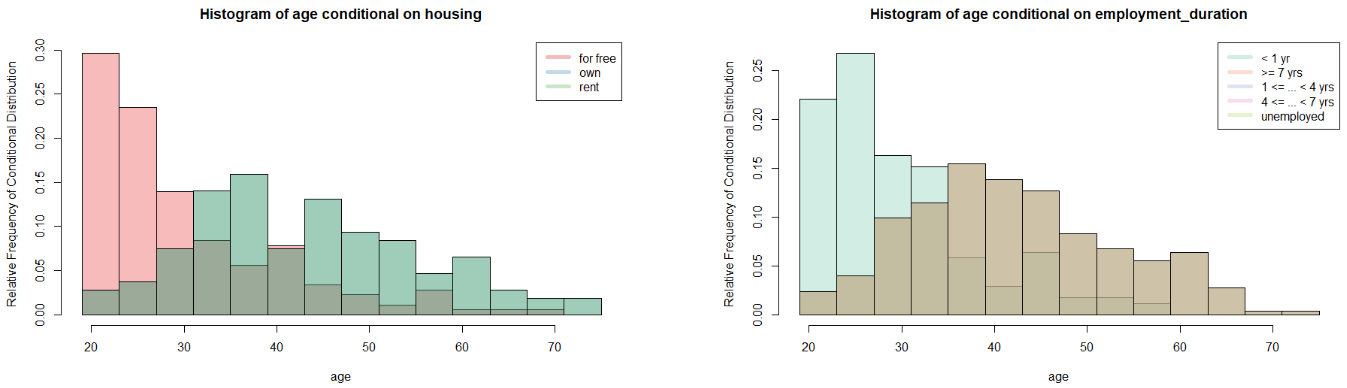

Figure 6: The probability of *age* given that we know *housing*, and the probability of *age* given that we know *employment_duration*.

These relationships are the strongest that can be observed when we condition our numerical variables on our categorical ones. The plots of all other variables produced overlapping histograms across the levels of the factor we are conditioning on. This implies that the qualitative variable in question does not impact the numerical one very much (at least this is what we can gather by eye anyway). Conditioning in the opposite direction was not possible for the scope of this report, a point we shall discuss later in our conclusion.

6 Data Quality Post-Analysis

After scrutinising the data throughout our exploratory analysis, some strange (and possibly erroneous) patterns began to surface. In this section, we’ll look to describe some evidence of this and briefly state what this might mean for the study itself. Figure 7 shows a relative frequency bar plot of the *job* variable conditional on the *employment_duration* variable. If we study it closely we can see that, given a randomly selected debtor who is unemployed, there was a 53% chance of them being a ‘manager/self-employed/highly qualified employee’, a 1.6% chance of them being an ‘unskilled‘ resident, and a 19% chance of them being a ‘skilled employee/official’. There is clearly an inconsistency in the data here and begs the question, **how many other records or variables are like this?** When the data was collected perhaps the methods involved were not completely watertight, the problem here is not even necessarily with the ordinal levels of *employment_duration* or *job*, it lies with the fact that in the ‘unemployed’ level of the *employment_duration* attribute, the blue bar should be set to a value of 1, and the other bars should be 0 (this is because all unemployed debtor’s **must** belong to the ‘unemployed/unskilled-non-resident’ level of the *job* variable by definition).

After analysing the relationships between **all** explanatory variables in section 5, this small discrepancy seems to be the only major outlier. This is good news as it means that perhaps our **entire** dataset is not effected by our recent observation and (at least since our sample is reasonably large) could still hold some statistical significance to the response variable, even **if** some records are erroneous. It doesn’t make sense to simply remove these erroneous records and move on as it could skew our predictions and we have no idea if all the attributes in these records are erroneous, they could provide perfectly accurate data of the response variable. However, in our model fitting we should be skeptical when including either of these variables in our final model.

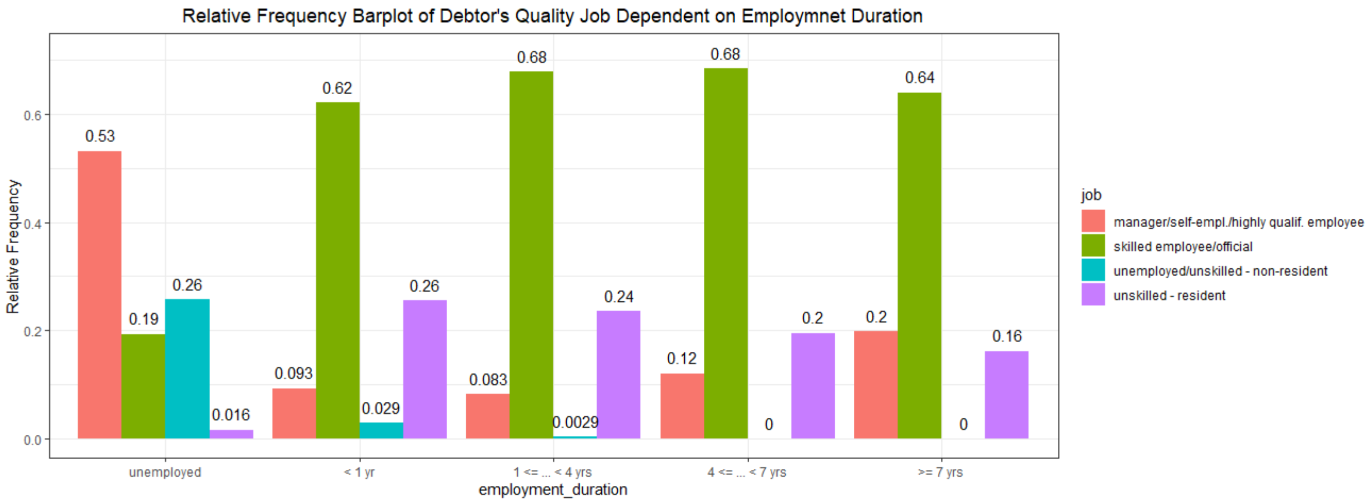

Figure 7: In addition to the issues regarding this plot, we also encountered a discrepancy in the *housing* and *property* variables in section 5.

7 Conclusion

7.1 Summary of Conclusions

Finally, let's assess what we've learnt about our data, the type of models that we could fit, and perhaps which attributes may not be suitable for analysis and why. From Figure 2, we saw that the majority of our variables directly impacted *credit_risk*, this is a good sign, it means we have many covariates which can lead to a change in a debtor's credit risk and thus we have many relationships to use for our model. However, *installment_rate, present_residence, number_credits, job* and *people_liable* were all shown to be poor variables to assess how *credit_risk* changes. Yet we concluded still that since there could be strong relationships between these insignificant variables and other **explanatory** variables (and thus they could be rather useful when mode fitting), we decided not to omit them entirely. Two outliers to this rule were the *personal_status_sex* and *people_liable* attributes. Omitting the first variable because it does not contain solid information about sex **or** marital status in isolation at all due to the fact that both traits were merged (as noted in section 4), and the second variable because it received a p-value of 1 in our hypothesis tests (making it exactly independent from *credit_risk* and thus, even through indirect relationships as stated earlier, it is practically useless).

In section 3, we considered a GLM to fit our data with, the data seemed to fit the criteria required for such an approach. On the other hand, we also mused on the idea of a decision tree. As such, we may require a thorough and rigorous evaluation of both methods after performing these analysis techniques to compare these approaches in detail. Throughout section 6, we explored the seemingly erroneous data that reared its head post-analysis. It was found that *job, employment_duration, housing* and *property* harbour strange relationships that **should not be possible**, and thus we should be careful when including them in any model that we fit.

7.2 Limitations of our Exploratory Analysis

Although we've delved into our data thoroughly, due to the scope of this report, there are still some limitations we can observe post-analysis:

- It may have been informative to investigate and visualise the information value (IV)contributed by our explanatory variables.
- The condition relationships in section 5.3 were only conducted in one direction, we did not test for [*categorical variable*] dependent on [*numerical variable*].
- In addition to this, the plots in section 5.3 showed many overlapping histograms (one for each level of the categorical variable). Thus, some of the histograms were difficult to see properly over others.

References

[1] U. Grömping, "South german credit data: Correcting a widely used data set," 2019.

[2] "Machine learning repository, https://archive.ics.uci.edu/ml/datasets/south+german+credit," 2019.