## Classification goals

The task is a binary classification problem. The goal is to predict Bad clients (target=1) based on application information provided.

## Metrics

We need to choose a metric that will show the rate of right predictions. Common metrics for Binary classification tasks are ROC AUC and Gini.
For the sake of the reducing losses (risk minimization), between two algorithms with close quality metrics the one that will have less False Negative rate will be chosen. This rate shows the number of Bad clients that were misclassified as Good and the greater False Negative rate can lead to future losses when loan is approved.

## Target

Apply mapping: Good:0, Bad:1
Result: Column with [0,1] values

## Missing values

Only one column 'job type' has missing values (for 234 clients).
To fill missing value new class was added (value=0)

## Binary categorical variables

Apply mapping with respect to correlation with Target
Result: Column with [0,1] values

## Categorical variables (more than 2 values)

One Hot Encoding was applied to deal with Categorical variables with more than 2 classes. Less frequent classes were dropped.

## Ways to study feature influence and importance

1. Comparing means grouped by target
   Comparing means of the columns grouped by target shows that some features have quite close mean values for “good” and “bad” clients and that is why they may not have strong predictive power. Among such feature are: “foreign”, “telephon”, “dependents”, “resident”.
2. Categorical-categorical association
   Statistics Chi squared, Cramers V, Crosstabs, confusion matrix
3. Visualization
   Plots for every column grouped by target:
   - Stacked Bar Plot (category-target)
   - Histogram
   - BoxPlot

## Numerical variables

There are 4 numerical variables out of 22 variables: ‘age’, ‘loan duration (m), ‘gross payments’ and ‘principal payments’.

### 'Gross payments' and 'Principal payments'

'Gross payments' and 'principal payments' have strong positive correlation (0.998) and we can drop one of this columns. As 'principal payments' has a slightly higher correlation with target we will keep this column.

There can be identified that among the clients with a large 'principal payments' there is a large percentage of “bad” target.

‘Age’

The distribution of the age is skewed to the left, most of the clients are of the age between 22 and 42. There are 6 clients with the age over 70, only one of them is treated as a “bad” client. Younger people tend to have higher ration of “bad” classification.

'loan duration (m)'

The longer the credit the higher the chances to qualify client as “Bad” one.

All 4 are skewed to the right, log-transformation was applied.

## Categorical variables

‘History (number of loans)’ demonstrates negative “correlation” with target.
For clients with 0 and 1 loans the percentage of “Bad” as clients with more loans know that it is worth to be a ‘responsible’ client to be able to get another loan.

## New Features

1. Binarisation. Find values that separate classes.
2. Combining classes to increase the number of clients in less frequent

## Model

Logistic Regression from the “sklearn” library.

## Validation

To train training and validate models was applied the following procedure:

1. Dataset was splitted into 2 parts (Train/Test ratio=0.3) with stratification to the Target class
2. As minor class “BAD” share is 30% in the dataset the oversampling was applied to the Trainset to balance classes. To balance classes function SMOTE from imblearn.over_sampling library was applied.
3. For the Oversampled Trainset KFold was split into folds (N=6).
4. N models were trained on N-1 fold.
   Validation: Predict on a left-off folds and calculate quality metrics. Mean of quality metrics was calculated.
   Validation (out of sample): Predict on Testset. Quality metrics calculated. Predictions for class and probability was stored
5. Aggregated prediction:
   a. Prediction of class – majority rule (mean>=0.5)
   b. Prediction of probability – mean function.

In a nutshell, the quality of the solution is measured on 2 kinds of sets:

1. Left-off folds (balanced data- Valid1)
2. Testset (unbalanced data, 30% of the original dataset – Valid2)

## Feature selection

Iterative procedure was applied to eliminate features. On each iteration, the feature that resulted in the largest increase of the quality if not applied was eliminated.
Quality was calculated as minimum from Score_1 and Score_2:

Score_1 = Mean of the Metric score on Left-off folds

Score_2 = Metric score on aggregated prediction

## Results

Train/test split=0.7/0.3
Shape_0: 83 (all in)
Shape: 71 (after feature selection)
Folds 6 Seed 101 Selection: 1
Feature selection based on Gini score.

**Validation scores on left-off folds (196 clients)**
------------min folds--------------
GINI 0.5735  Roc AUC 0.7868  FN 15  FP 22
------------mean folds--------------
GINI 0.6408  Roc AUC 0.8204  FN 18.7  FP 23.3
------------std folds--------------
GINI 0.0432  Roc AUC 0.0216  FN 2.9  FP 1.8

**Validation scores on Testset (300 clients)**
------------min folds--------------
GINI 0.6232  Roc AUC 0.8116  FN 21  FP 45
------------mean folds--------------
GINI 0.6316  Roc AUC 0.8158  FN 23.7  FP 48.0
------------std folds--------------
GINI 0.0074  Roc AUC 0.0037  FN 1.8  FP 2.2

**Aggregated prediction  on Test_set  (300 clients)**
Confusion matrix

| | 0 | 1 |
|---|---|---|
| 0 | 163 | 47 |
| 1 | 23 | 67 |

Class 1 (BAD):  114 Class 0 (Good):  186
Misclassified BAD: 23 (7.667%) !!!
Misclassified GOOD: 47 (15.667%)
Accuracy 76.67%
------
GINI: 0.6386 Roc AUC 0.8193 FN 23 FP 47

## Features

**Numerical and “ordinal”**
1. checking
2. purpose of loan
3. savings
4. installp
5. marital
6. co-applicant status
7. resident
8. property
9. other

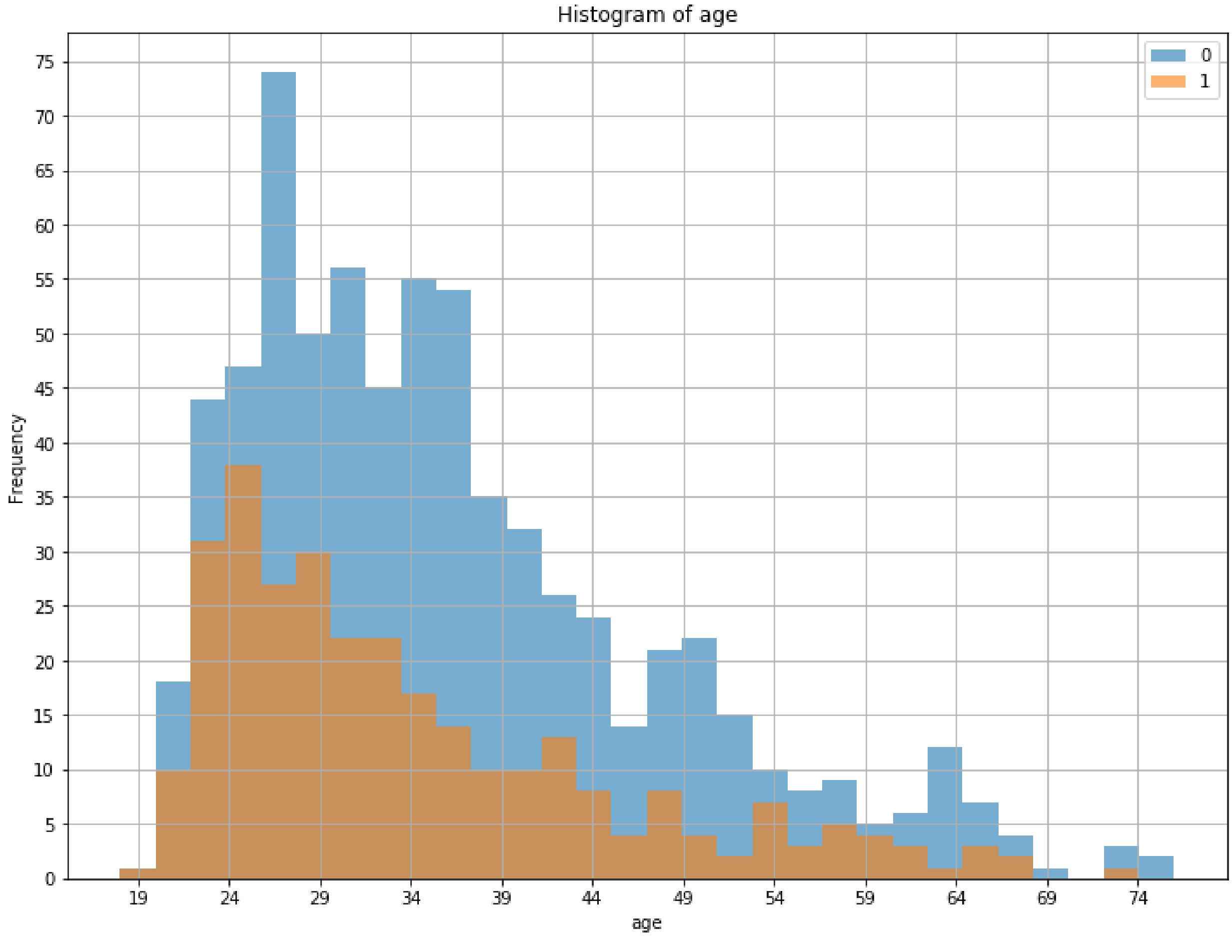
10. dependents
11. telephon
12. foreign
13. job type_m
14. other_bin
15. job_bin
16. housing_bin
17. job type_mod
18. marital_mod
19. age_l
20. history (number of loans)
21. principal payments_l
22. loan duration (m)_l
23. checking_bin
24. marital_bin
25. history (number of loans)_mod
26. savings_bin
27. principal payments_bin
28. loan duration (m)_bin
29. property_bin

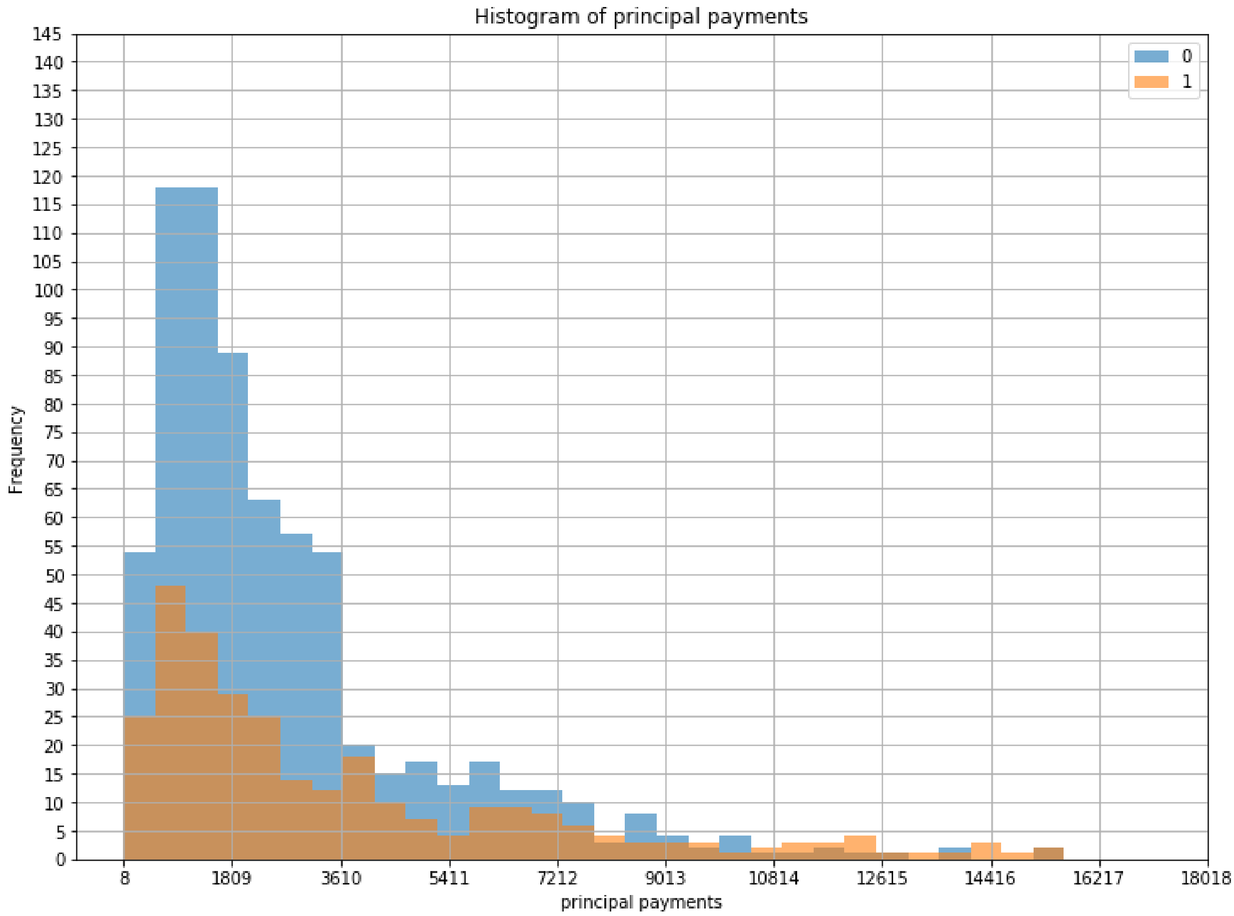
**Categorical after One Hot Encoding**

1. OHE_checking_4
2. OHE_checking_1
3. OHE_history (number of loans)_2
4. OHE_history (number of loans)_4
5. OHE_history (number of loans)_3
6. OHE_purpose of loan_0
7. OHE_purpose of loan_2
8. OHE_purpose of loan_1
9. OHE_purpose of loan_9
10. OHE_purpose of loan_6
11. OHE_purpose of loan_5
12. OHE_savings_1
13. OHE_savings_5
14. OHE_savings_2
15. OHE_savings_3
16. OHE_employed_5
17. OHE_employed_4
18. OHE_employed_2
19. OHE_installp_4
20. OHE_installp_2
21. OHE_marital_3
22. OHE_marital_2
23. OHE_marital_4
24. OHE_co-applicant status_1
25. OHE_co-applicant status_3
26. OHE_resident_4
27. OHE_resident_2
28. OHE_resident_3
29. OHE_property_3
30. OHE_property_1

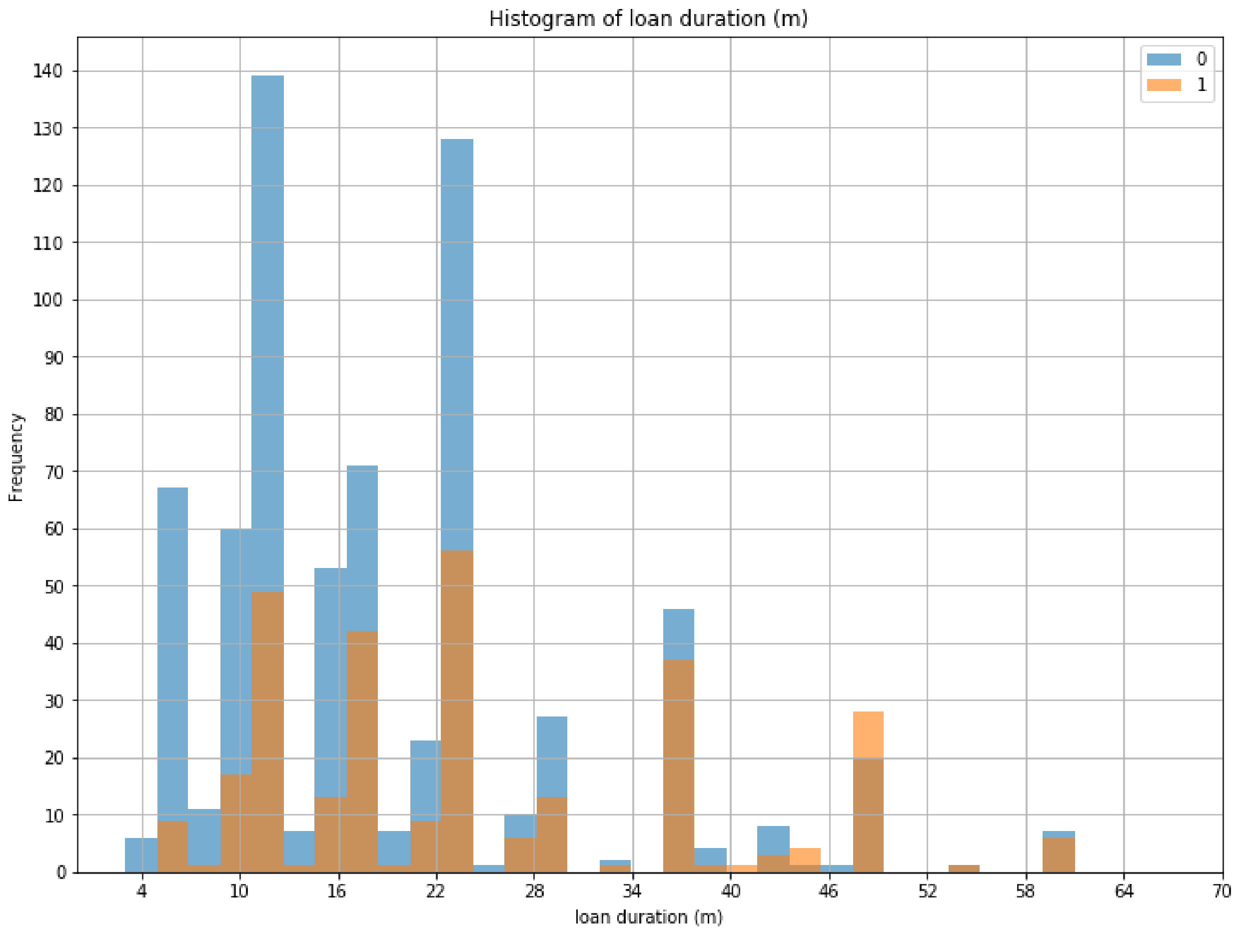
31. OHE_property_2
32. OHE_other_3
33. OHE_other_1
34. OHE_housing_1
35. OHE_exist credit bureau data_1
36. OHE_exist credit bureau data_2
37. OHE_exist credit bureau data_3
38. OHE_job_3
39. OHE_job_2
40. OHE_job_4
41. OHE_job type_3.0
42. OHE_job type_4.0

# Visualisation

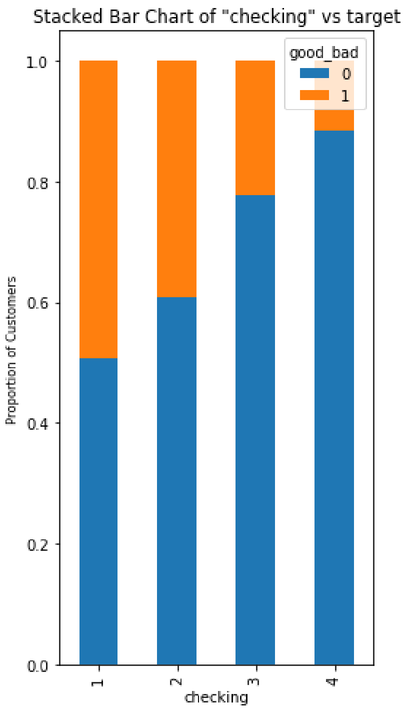
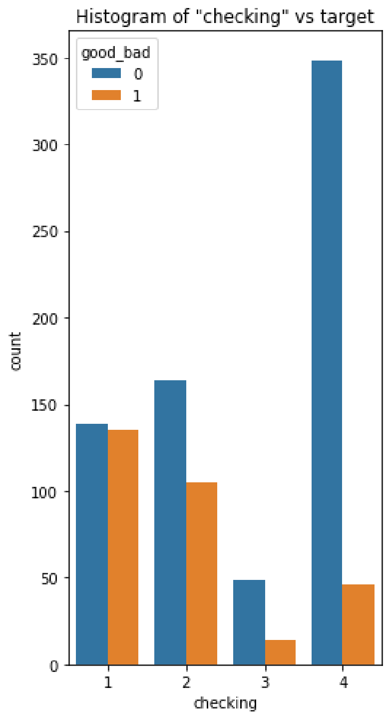
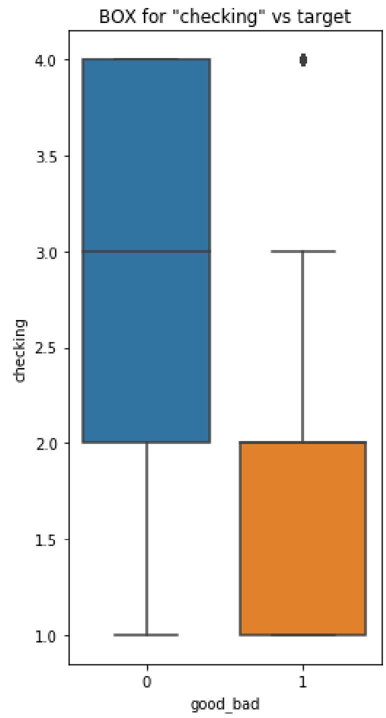
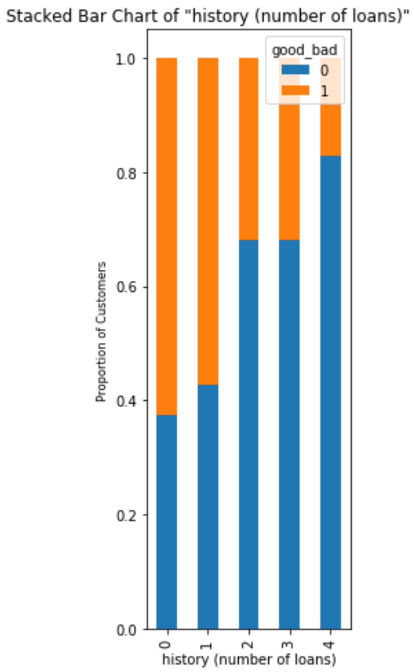
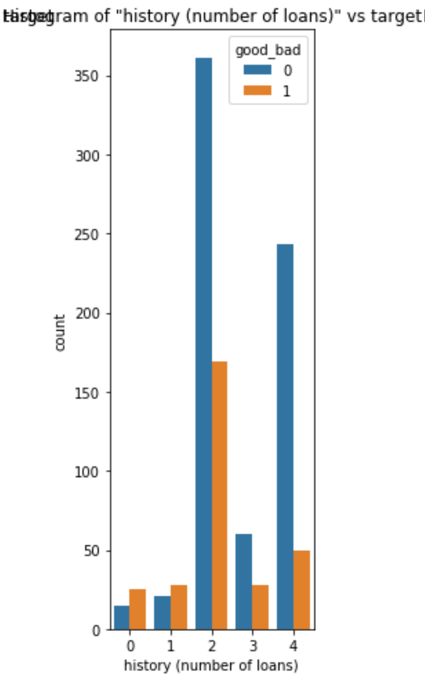
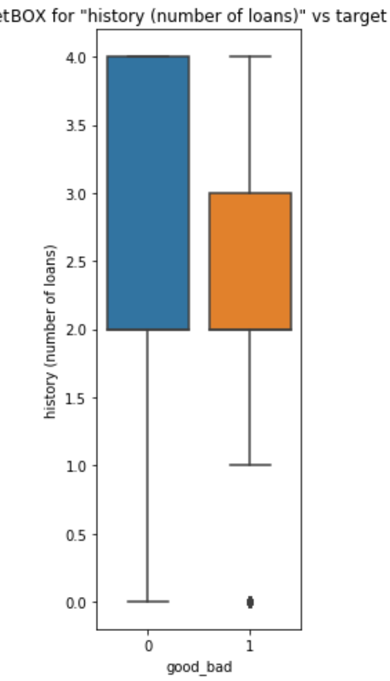


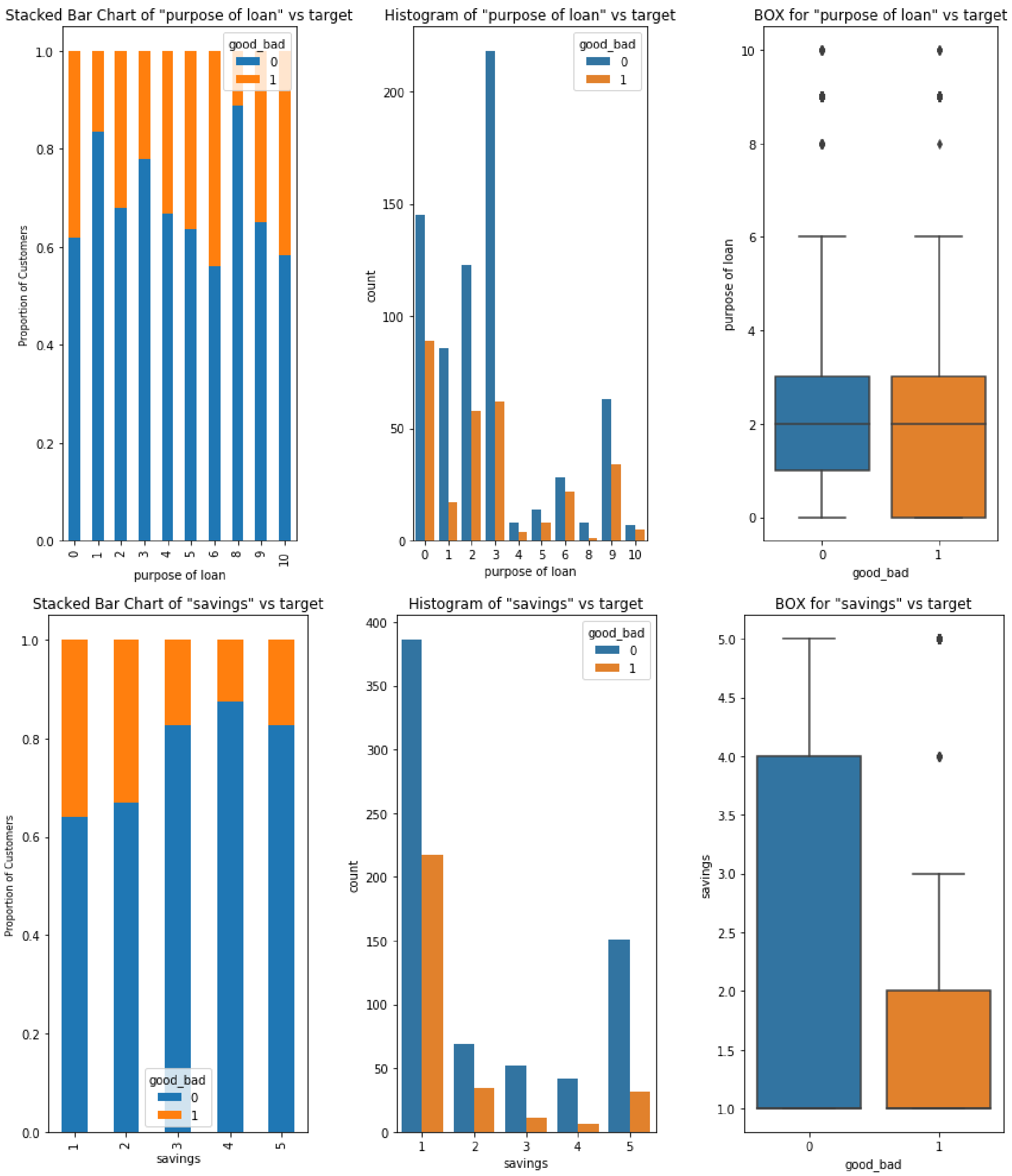


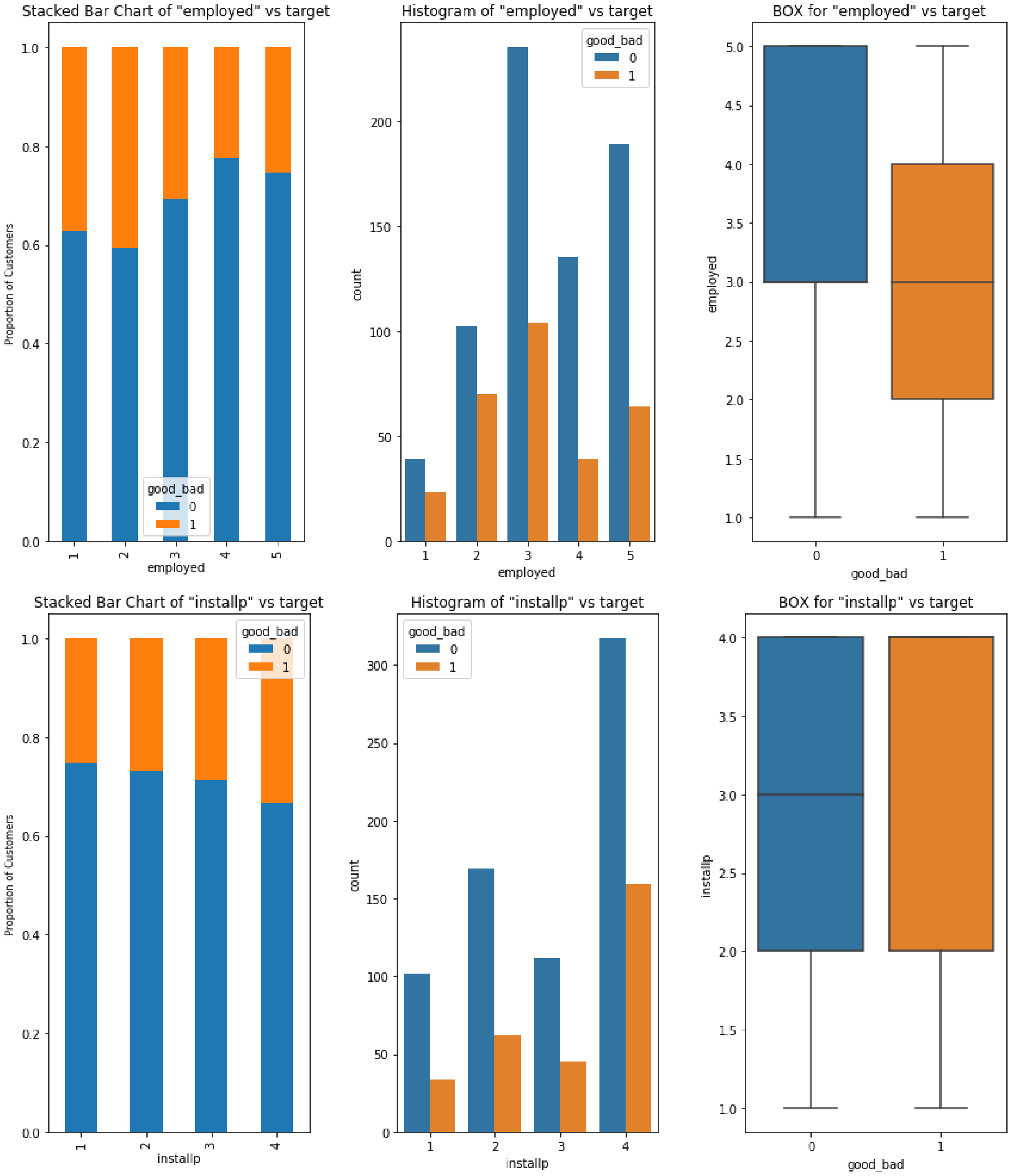
Histogram of principal payments
Frequency
principal payments
0
1
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
8
1809
3610
5411
7212
9013
10814
12615
14416
16217
18018

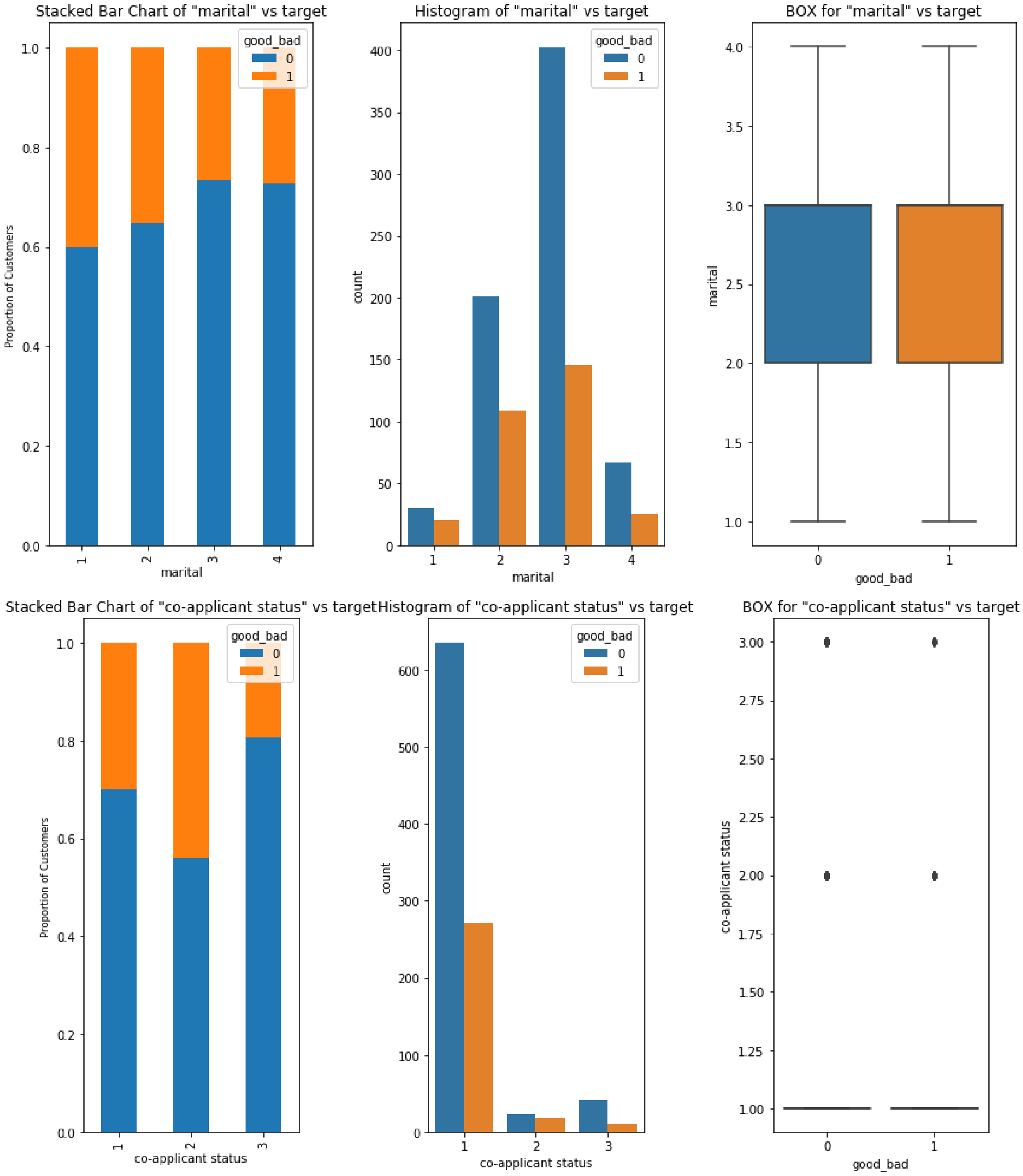
Histogram of loan duration (m)
Frequency
loan duration (m)
0
1
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
4
10
16
22
28
34
40
46
52
58
64
70

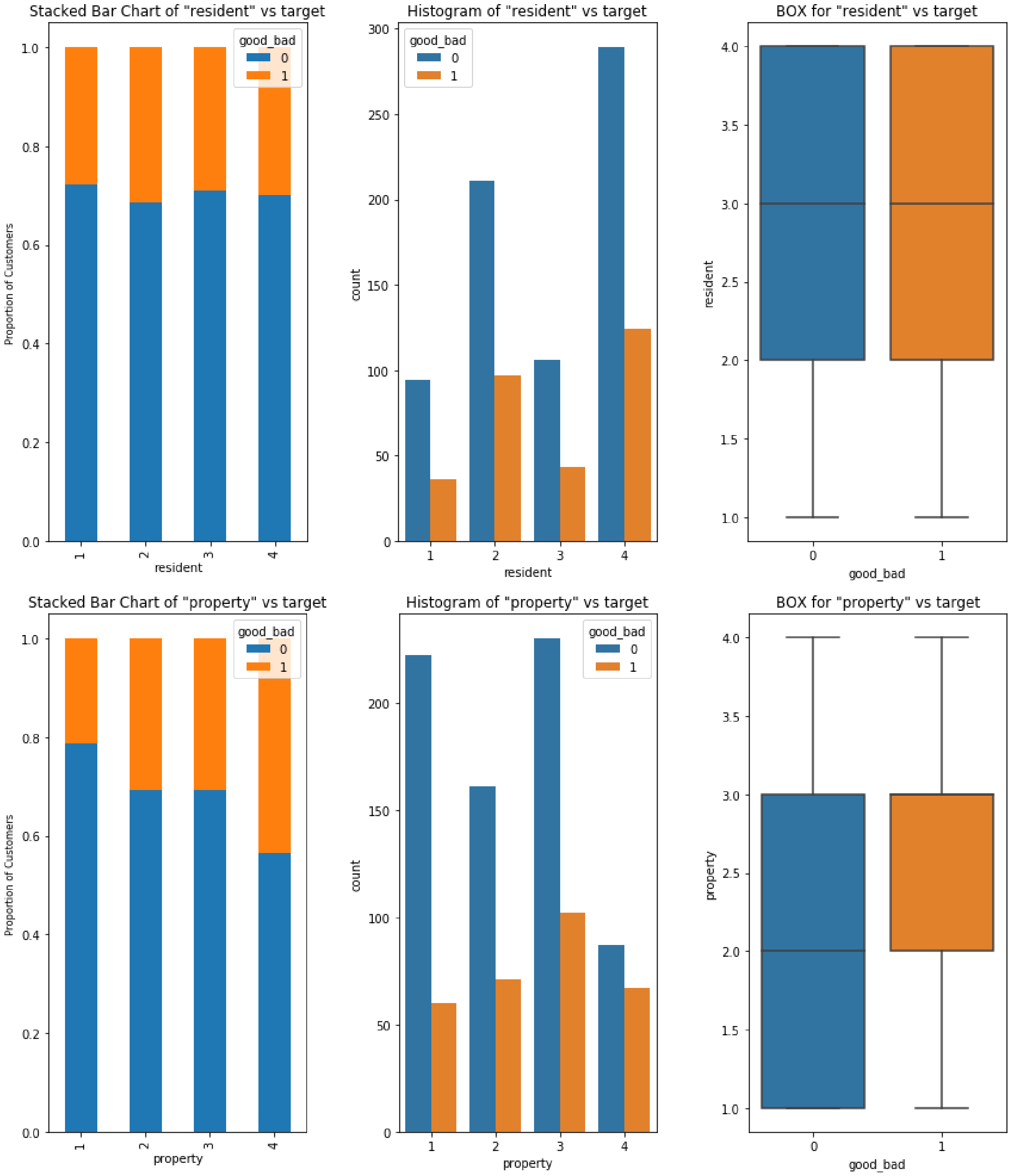

Stacked Bar Chart of "checking" vs target
Proportion of Customers
checking
good_bad
0
1

Histogram of "checking" vs target
count
checking
good_bad
0
1

BOX for "checking" vs target
checking
good_bad

Stacked Bar Chart of "history (number of loans)" vs target
Proportion of Customers
history (number of loans)
good_bad
0
1

Histogram of "history (number of loans)" vs target
count
history (number of loans)
good_bad
0
1

BOX for "history (number of loans)" vs target
history (number of loans)
good_bad


Stacked Bar Chart of "purpose of loan" vs target
Proportion of Customers
purpose of loan
good_bad
0
1
Histogram of "purpose of loan" vs target
count
purpose of loan
BOX for "purpose of loan" vs target
purpose of loan
good_bad
Stacked Bar Chart of "savings" vs target
Proportion of Customers
savings
Histogram of "savings" vs target
count
savings
BOX for "savings" vs target
savings
good_bad

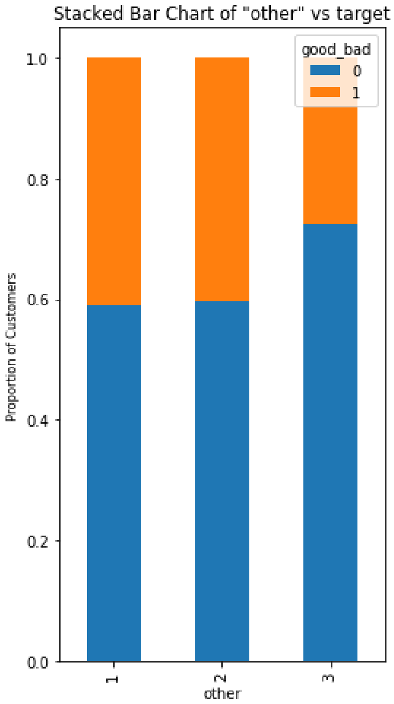


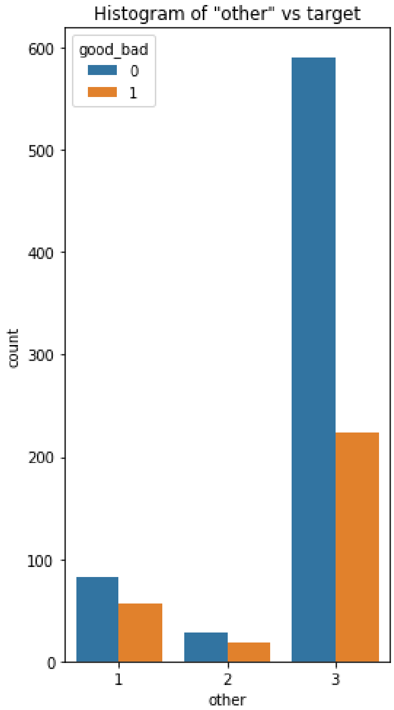
Stacked Bar Chart of "employed" vs target
Proportion of Customers
employed
good_bad
0
1
Histogram of "employed" vs target
count
employed
BOX for "employed" vs target
employed
good_bad
Stacked Bar Chart of "installp" vs target
Proportion of Customers
installp
Histogram of "installp" vs target
count
installp
BOX for "installp" vs target
installp
good_bad

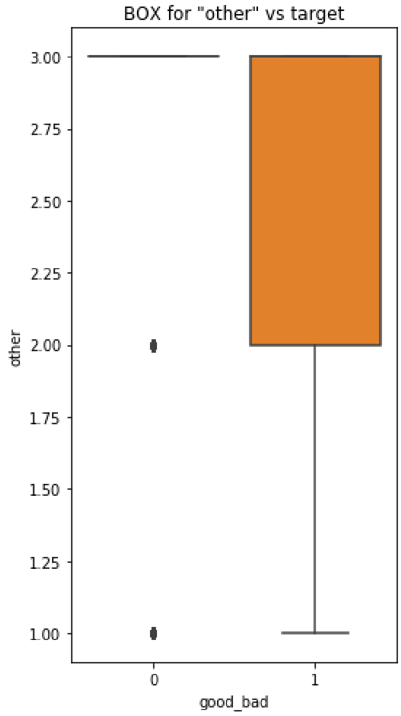

Stacked Bar Chart of "marital" vs target
Proportion of Customers
marital
good_bad
0
1
Histogram of "marital" vs target
count
marital
BOX for "marital" vs target
marital
good_bad
Stacked Bar Chart of "co-applicant status" vs target
Proportion of Customers
co-applicant status
Histogram of "co-applicant status" vs target
count
co-applicant status
BOX for "co-applicant status" vs target
co-applicant status
good_bad

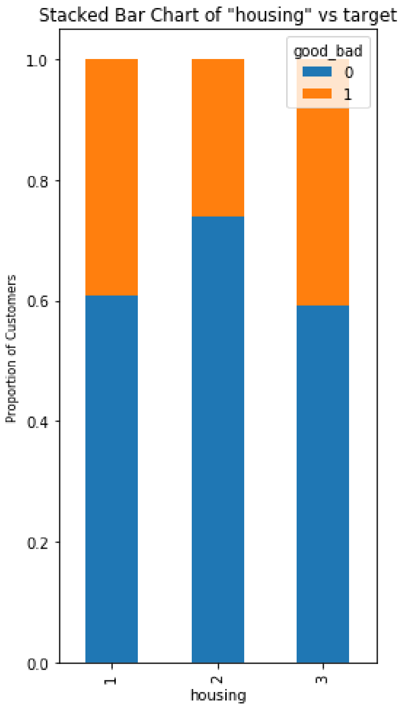


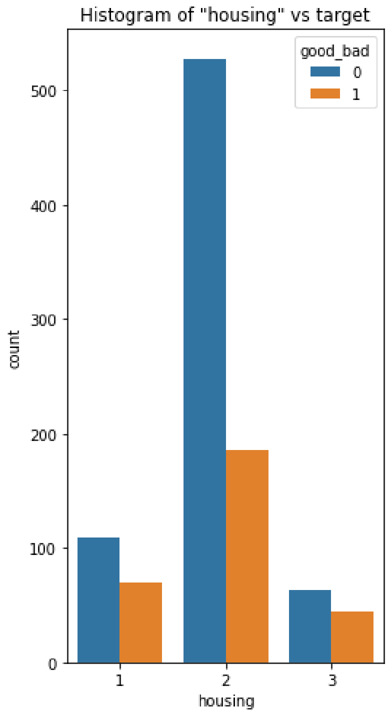
Stacked Bar Chart of "resident" vs target
Proportion of Customers
resident
good_bad
0
1
Histogram of "resident" vs target
count
resident
BOX for "resident" vs target
resident
good_bad
Stacked Bar Chart of "property" vs target
Proportion of Customers
property
Histogram of "property" vs target
count
property
BOX for "property" vs target
property
good_bad

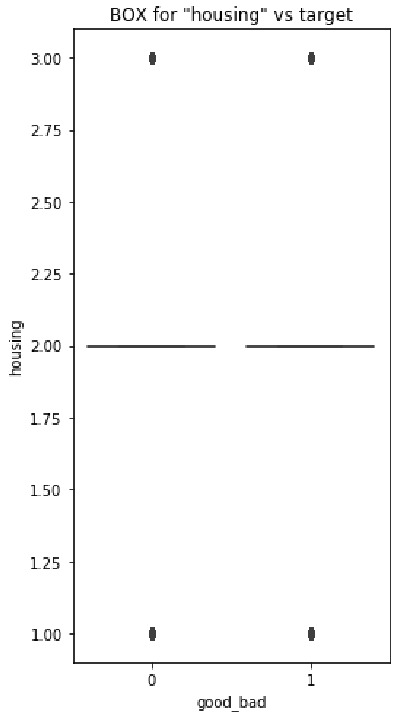

Stacked Bar Chart of "other" vs target
good_bad
0
1
Proportion of Customers
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
other

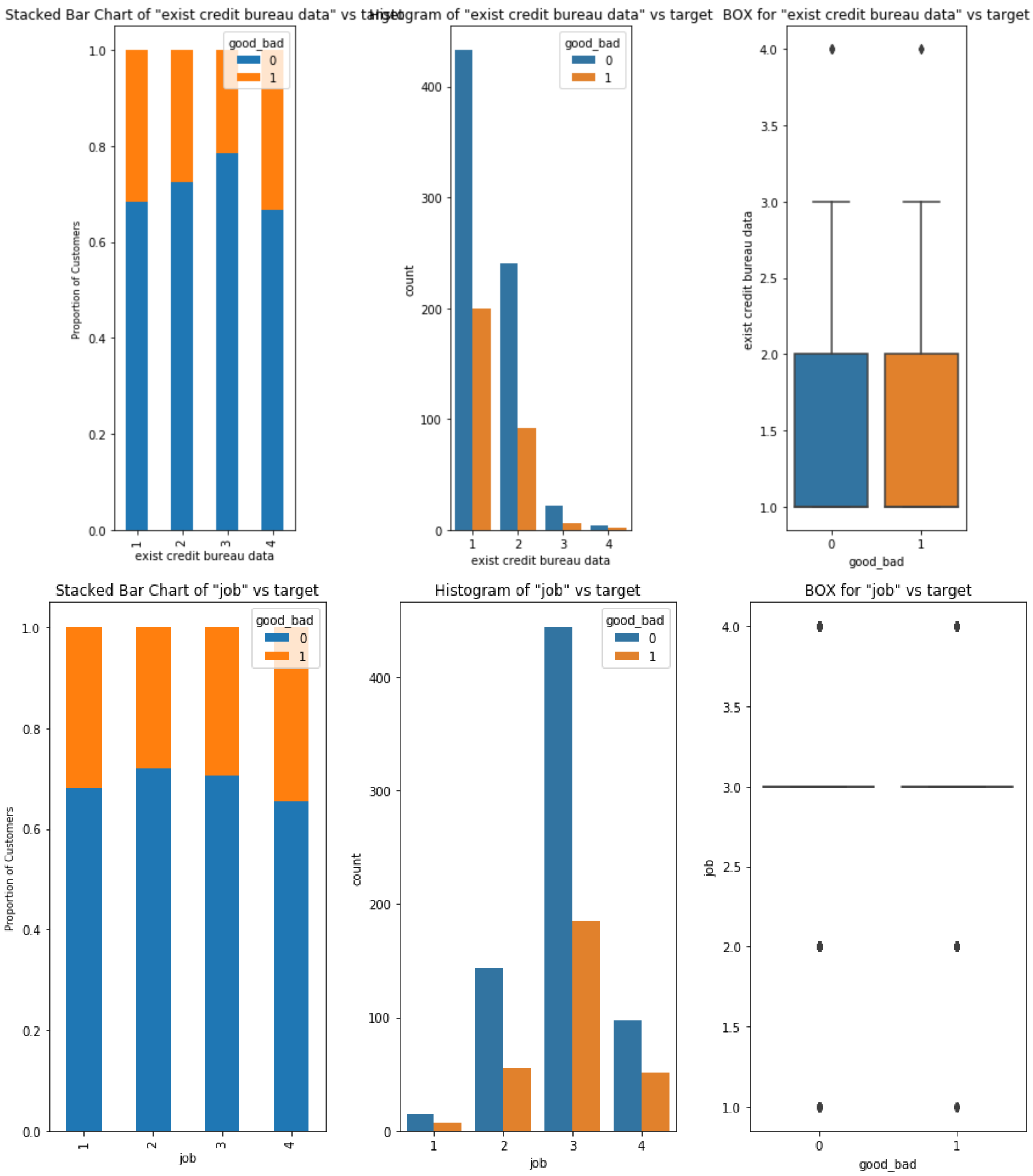

Histogram of "other" vs target
good_bad
0
1
count
0
100
200
300
400
500
600
1
2
3
other


BOX for "other" vs target
other
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
0
1
good_bad

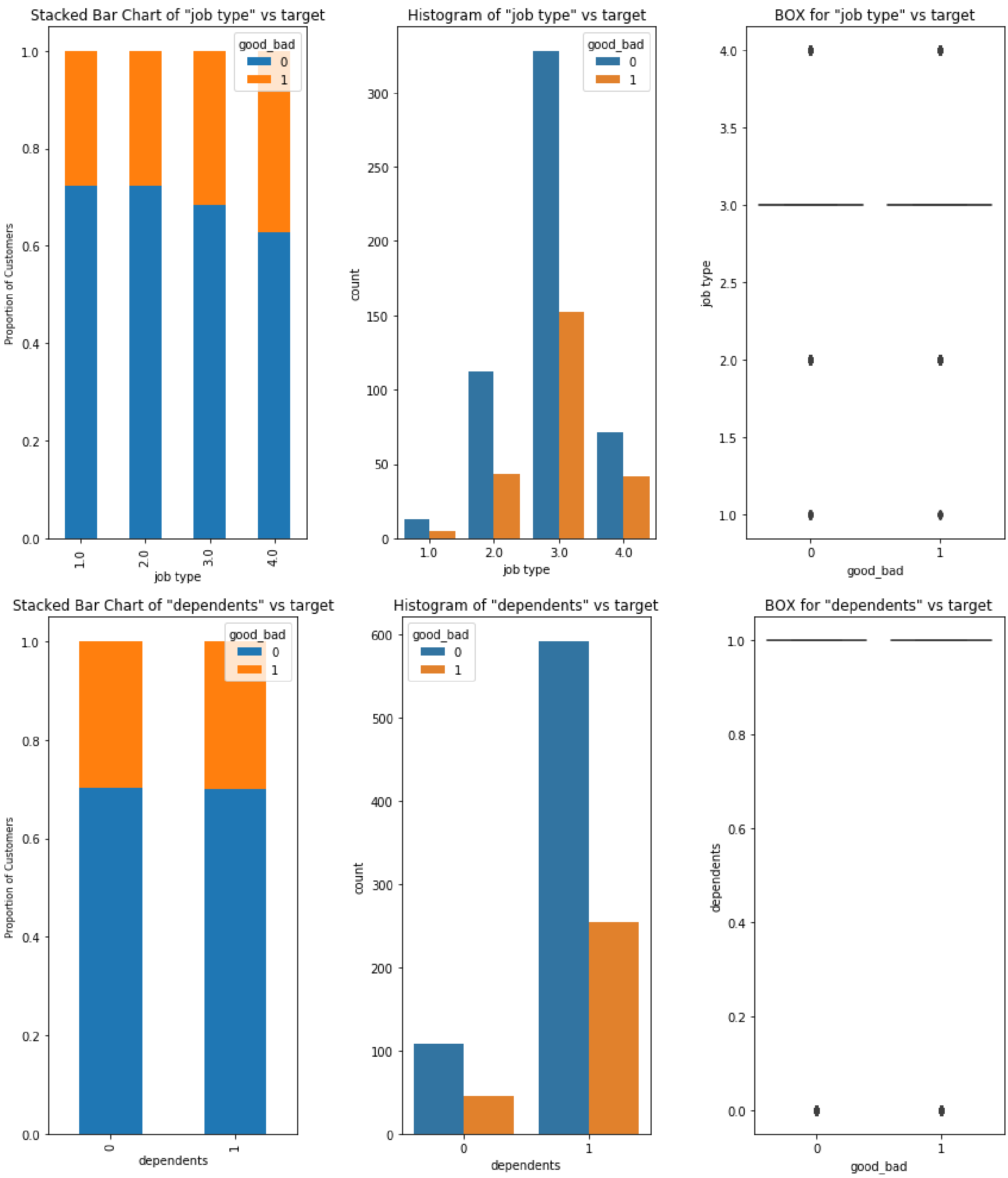

Stacked Bar Chart of "housing" vs target
good_bad
0
1
Proportion of Customers
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
housing

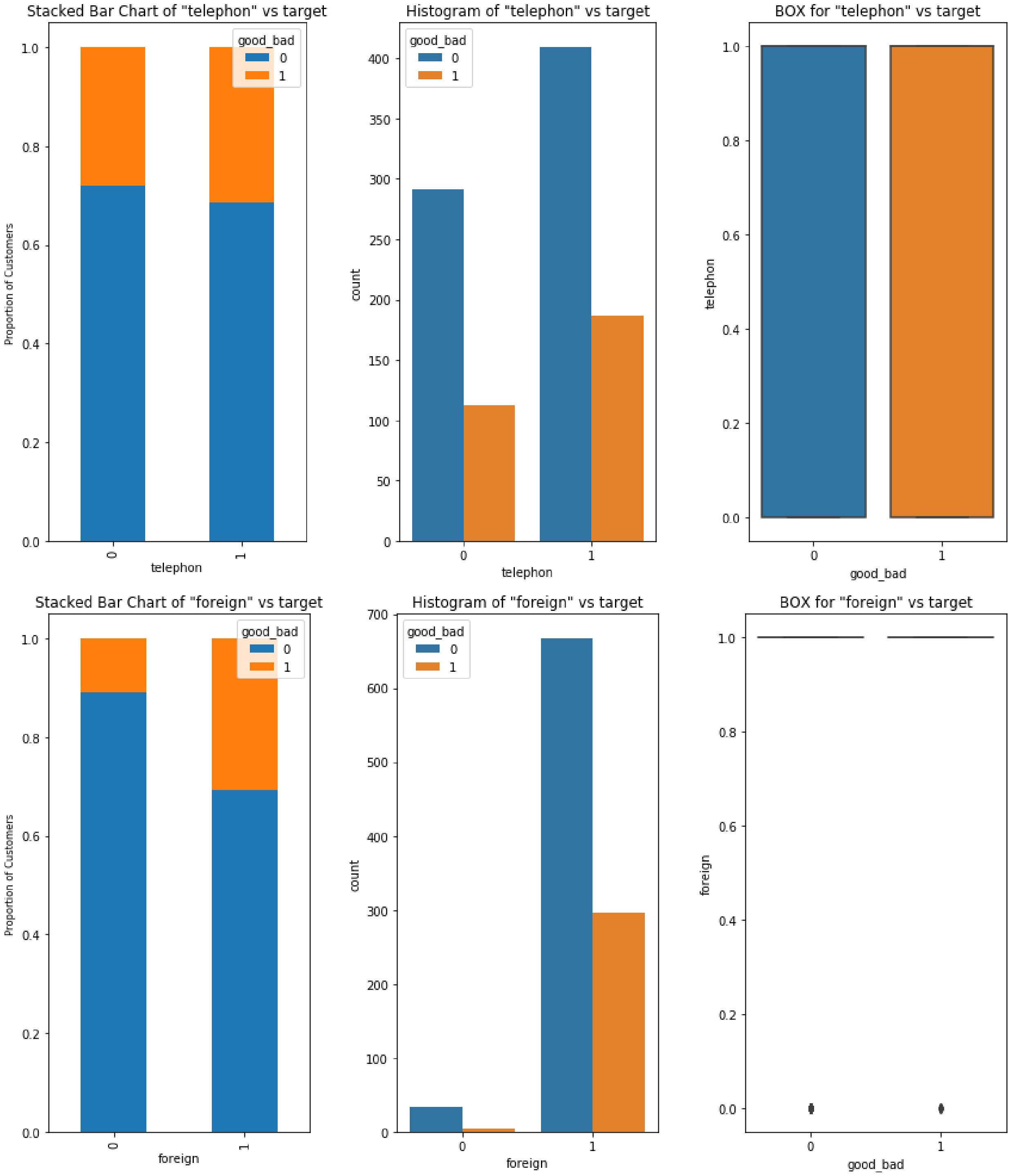

Histogram of "housing" vs target
good_bad
0
1
count
0
100
200
300
400
500
1
2
3
housing


BOX for "housing" vs target
housing
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
0
1
good_bad



Stacked Bar Chart of "exist credit bureau data" vs target
Proportion of Customers
exist credit bureau data
good_bad
0
1
Histogram of "exist credit bureau data" vs target
count
exist credit bureau data
good_bad
0
1
BOX for "exist credit bureau data" vs target
exist credit bureau data
good_bad
Stacked Bar Chart of "job" vs target
Proportion of Customers
job
good_bad
0
1
Histogram of "job" vs target
count
job
good_bad
0
1
BOX for "job" vs target
job
good_bad


Stacked Bar Chart of "job type" vs target
Proportion of Customers
job type
good_bad
0
1
Histogram of "job type" vs target
count
job type
BOX for "job type" vs target
job type
good_bad
Stacked Bar Chart of "dependents" vs target
Proportion of Customers
dependents
Histogram of "dependents" vs target
count
dependents
BOX for "dependents" vs target
dependents
good_bad


Stacked Bar Chart of "telephon" vs target
Proportion of Customers
telephon
good_bad
Histogram of "telephon" vs target
count
telephon
BOX for "telephon" vs target
telephon
good_bad
Stacked Bar Chart of "foreign" vs target
Proportion of Customers
foreign
Histogram of "foreign" vs target
count
foreign
BOX for "foreign" vs target
foreign
good_bad