Peer Landing Project

Stage 2 - Data Understanding/Data Preparation
Group C

In this stage of the project, we performed EDA, cleaned the data, filtered rows (loans) and columns (attributes), got a better understanding of the attributes that might have prediction power to our model, and calculated the realized return we believe is true. After this stage, we will start examining the different models.

The steps we took:

1. **Binding** the quarters into 2 tables, one for 2018 and one for 2019.

2. **Cleaning the data by rows (e.g attributes) :**

   a.
   Loan Status - There are 8 categories (7+NA) and most of the loans are Charged off, Fully Paid, or Current. We decided to keep only loans that have ended - Charged off and Fully Paid since for ongoing loans we can't know what will happen in the future so we can't rely on these loans for our model.

   b.
   Loan Term - There are 2 options for this attribute, 36 months and 60 months, where around 80% of the loans are for 36 months. we decided to keep only loans for 36 months for two reasons :

   - the loans originated in 2016 therefore loans for 60 months are still current in our data
   - The population behavior can vary between the two loan terms so even loans that have ended (e.g. prepaid) cannot be trusted.

   c.
   Application Term- that column indicates whether the loan is an individual application or a joint application with two co-borrowers. We decided to keep only the loans of individual applications since the number of co-borrowers is negligent and the population behavior might vary.

3. **Dealing with N/A**

   We checked the fraction of N/A for every column and have decided on a threshold of 40%, which means any attribute with more than 40% of N/A will be dropped. The threshold was decided according to the first "big gap" we noticed in the NA percentage. The remaining columns containing missing values will be dealt with in the modeling stage. Using this criterion, we removed 44 columns (the columns are specified in appendix 1). When we will move forward to the modeling stage, we will use a more elaborate way to replace the missing values such as KNN for categorical attributes and Linear regression for numeric attributes.

4. **Creating realized return column**

   To do so we parsed the relevant date columns and checked which columns are contained in the total payment column. it seems that the "Collection recovery fee" is not in the total payment so we added it to our calculation.

We will make a few assumptions to calculate the expected return:

- All payments are equal, each of $\frac{p}{m}$, are made in each of the m periods
- Each payment is immediately reinvested at a rate of i (e.g. 2%)

Define:

- f - the full amount invested in a loan
- p - full amount recovered from the loan (payments + recoveries)
- t - nominal duration of the loan in months
- m - actual duration of the loan – from first to last payments
- i - interest reinvested rate

$$\frac{12}{T}\cdot\frac{1}{f}\left\{\left[\frac{p}{m}\left(\frac{1-(1+i)^{m}}{1-(1+i)}\right)\right](1+i)^{T-m}-f\right\}$$

We will sum up the geometric series of loan repayments, subtract the original loan and divide by the original loan amount. To get the expected annual return we will multiply by 12/t.

5. Cleaning the data by columns

a. Unique value
We looked for columns that have the same value in more than 90% of the loans, which means that this column won't have any power for prediction, in this stage, we dropped 17 columns (See Appendix 1.1).

b. Leakage
We compared the snapshots of 2018 and 2019 to discover changes between the columns' values. This is an important step since changed values might indicate attribute leakage, hence, the values in the columns are not necessarily the values submitted during the time the loan was taken and we can't "trust" them. These attributes might be problematic for our prediction which occurs at the time of a loan. We carefully examined their description and removed changed and irrelevant attributes. We kept "total_pymnt" and "collection_recovery_fee", since we used them to calculate realized return and we want to check their correlation with all features. In This stage, we dropped 9 columns (See Appendix 1).

c. Irrelevant
After filtering columns by the code, we went through the attributes manually and checked for their relevance to our model - we've decided to drop 24 columns (See Appendix 1.1).

From this point there is no need for the 2018 snapshot of the loans, we will keep cleaning and exploring the 2019 version.

6. Correlation:

We started by checking the correlation (spearman and Pearson) of attributes with our main attribute - realized return and the attributes used to calculate it - total_pymnt, collection_recovery_fee, and funded_amnt. For better visualization, we split our data into 3 sub-matrices and produced a heatmap for each one with our relevant attributes (See Appendix 2 - Correlation Graphs - 2.4+2.5).

It is hard to make decisions regarding what variables to drop based on the plots, however, it seems that no attribute has a strong linear correlation with realized return, That might suggest we should use a non-linear model and therefore we will not drop attributes based on that (yet). There are few attributes with better spearman correlation with the realized return but also quite small

From the heatmaps, we saw that some attributes have strong linear correlation between each other (not the “main” attributes), this might suggest that one of every pair is redundant. However, we will not drop them now, instead we will create a table with every feachers combination and use it during feature selection in the modeling stage. Some of the attributes will be dropped based on that or based on a feature selection metric (e.g. forward selection), alternately we will consider creating new variables based on attributes with strong correlation to maximize our benefit from them.

After checking both linear and non-linear correlation of all attributes with our main features, we decided to remove all attributes that have less than 0.07 absolute correlation, both Pearson and spearman.
We set 0.07 as the threshold because we saw that attributes that have a higher correlation might be useful for our model based on their description

In this stage, we dropped 7 columns (See Appendix 1.1).

7. EDA

Appendix 2.7 - We can see that most of the loans have realized returns between 0.01 to 0.1. The loans with negative returns distribute almost equally approximately between -0.2 to -0.23.

Appendix 2.8 - It seems that all Fully Paid loans have positive returns which increase with the grade. We can see that are charged off loans with a positive return in every grade.

Appendix 2.9 - All possible purposes have approximately the same mean return besides wedding with a much higher return. It's unclear what is the reason for that, we will look into it further on.

In total - at this stage we removed 101 attributs.
We added 2 attributes - duration and realized return and therefore we have 51 columns for now (Appendix 1.2)

Appendix

Appendix 1:

- 1.1 Columns dropped

Column name	**Reason**
annual_inc_joint desc dti_joint member_id mths_since_last_delinq mths_since_last_major_derog mths_since_last_record mths_since_recent_bc_dlq mths_since_recent_revol_delinq next_pymnt_d verified_status_joint revol_bal_joint sec_app_fico_range_low sec_app_fico_range_high sec_app_earliest_cr_line sec_app_inq_last_6mths sec_app_mort_acc sec_app_open_acc sec_app_revol_util sec_app_open_act_il sec_app_num_rev_accts sec_app_chargeoff_within_12_mths sec_app_collections_12_mths_ex_med sec_app_mths_since_last_major_derog hardship_type hardship_reason hardship_status deferral_term hardship_amount hardship_start_date hardship_end_date payment_plan_start_date hardship_length hardship_dpd hardship_loan_status orig_projected_additional_accrued_interest hardship_payoff_balance_amount hardship_last_payment_amount debt_settlement_flag_date settlement_status settlement_date settlement_amount settlement_percentage settlement_term	NA -44

pymnt_plan term out_prncp out_prncp_inv total_rec_late_fee collections_12_mths_ex_med policy_code application_type acc_now_delinq chargeoff_within_12_mths delinq_amnt num_tl_120dpd_2m num_tl_30dpd num_tl_90g_dpd_24m tax_liens hardship_flag debt_settlement_flag	Unique - 17
recoveries last_pymnt_amnt last_pymnt_d total_rec_prncp total_rec_int last_fico_range_high last_fico_range_low total_pymnt_inv last_credit_pull_d	Leakage - 9
acc_open_past_24mt, earliest_cr_line funded_amnt_inv inq_last_12m issue_d mo_sin_old_il_acct mo_sin_old_rev_tl_op mo_sin_rcnt_rev_tl_op mo_sin_rcnt_tl mths_since_rcnt_il mths_since_recent_bc mths_since_recent_inq num_actv_rev_tl num_tl_op_past_12m open_acc_6m open_il_12m open_il_24m open_rv_12m open_rv_24m title total_acc url zip_code inq_last_6mths	Irrelevant -24

delinq_2yrs pub_rec open_act_il inq_fi total_cu_t num_accts_ever_120_pdl num_il_tl	Correlation -7

- 1.2 – Columns kept:

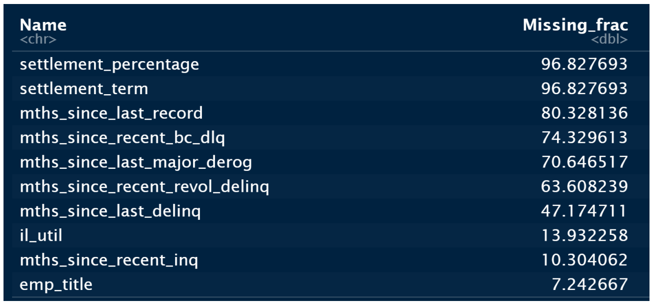
id
loan_amnt
funded_amnt
int_rate
installment
grade
sub_grade
emp_title
emp_length
home_ownership
annual_inc
verification_status
loan_status
purpose
addr_state
dti
fico_range_low
fico_range_high
open_acc
revol_bal
revol_util
initial_list_status
total_pymnt
collection_recovery_fee
tot_coll_amt
tot_cur_bal
total_bal_il
il_util
max_bal_bc
all_util
total_rev_hi_lim
avg_cur_bal
bc_open_to_buy
bc_util
mort_acc
num_actv_bc_tl
num_bc_sats

num_bc_tl
num_op_rev_tl
num_rev_accts
num_rev_tl_bal_gt_0

num_sats
pct_tl_nvr_dlq
percent_bc_gt_75
pub_rec_bankruptcies
tot_hi_cred_lim
total_bal_ex_mort
total_bc_limit
total_il_high_credit_limit
duration
realized_return

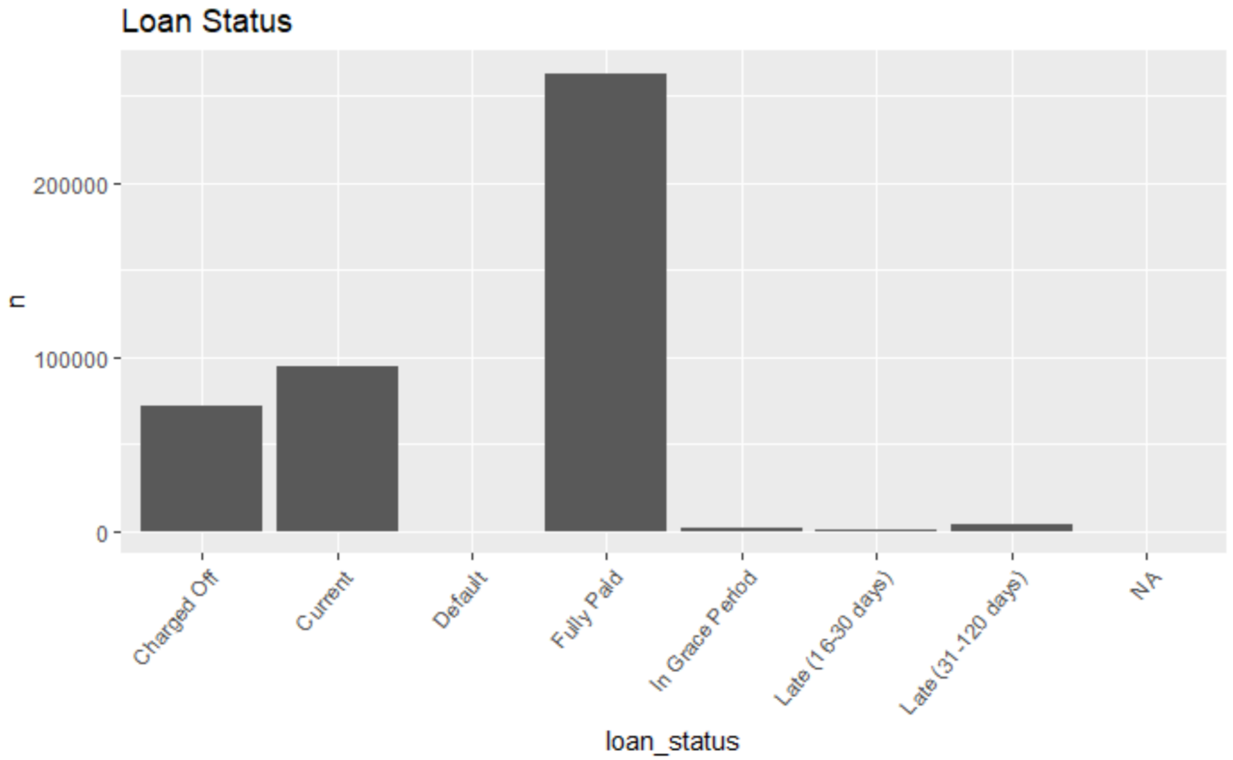
Appendix 2 - Graphs & visual examples

- 2.1 The “big step” of NA fraction

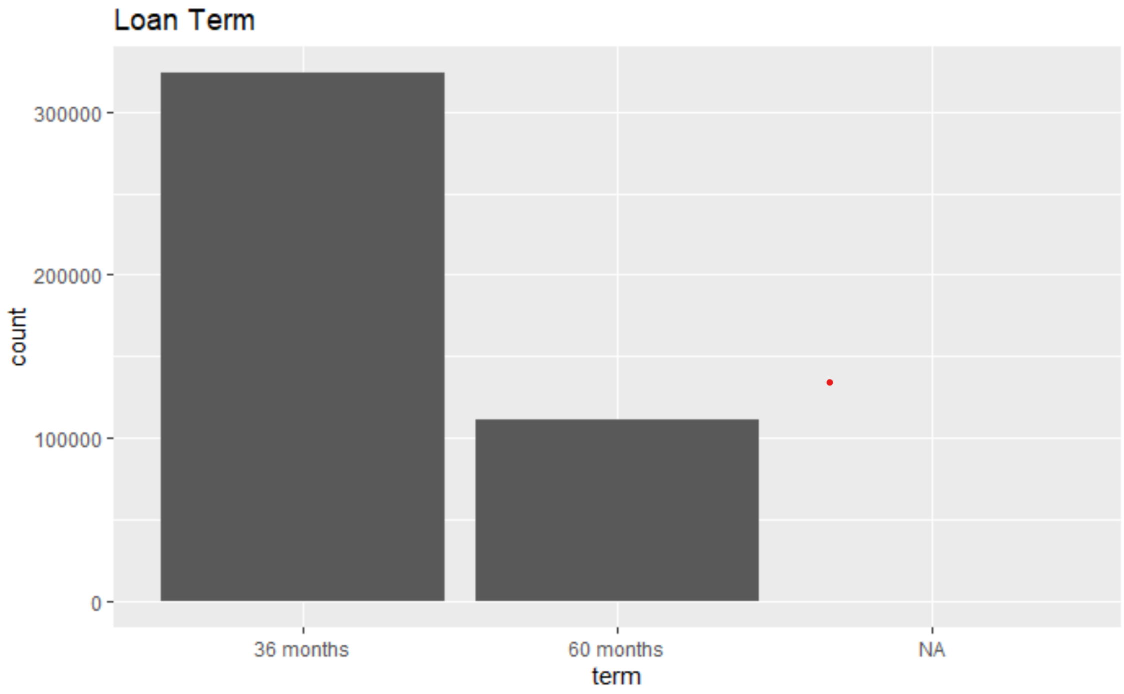


Name <chr>	Missing_frac <dbl>
settlement_percentage	96.827693
settlement_term	96.827693
mths_since_last_record	80.328136
mths_since_recent_bc_dlq	74.329613
mths_since_last_major_derog	70.646517
mths_since_recent_revol_delinq	63.608239
mths_since_last_delinq	47.174711
il_util	13.932258
mths_since_recent_inq	10.304062
emp_title	7.242667

- 2.2 Distribution of loan status

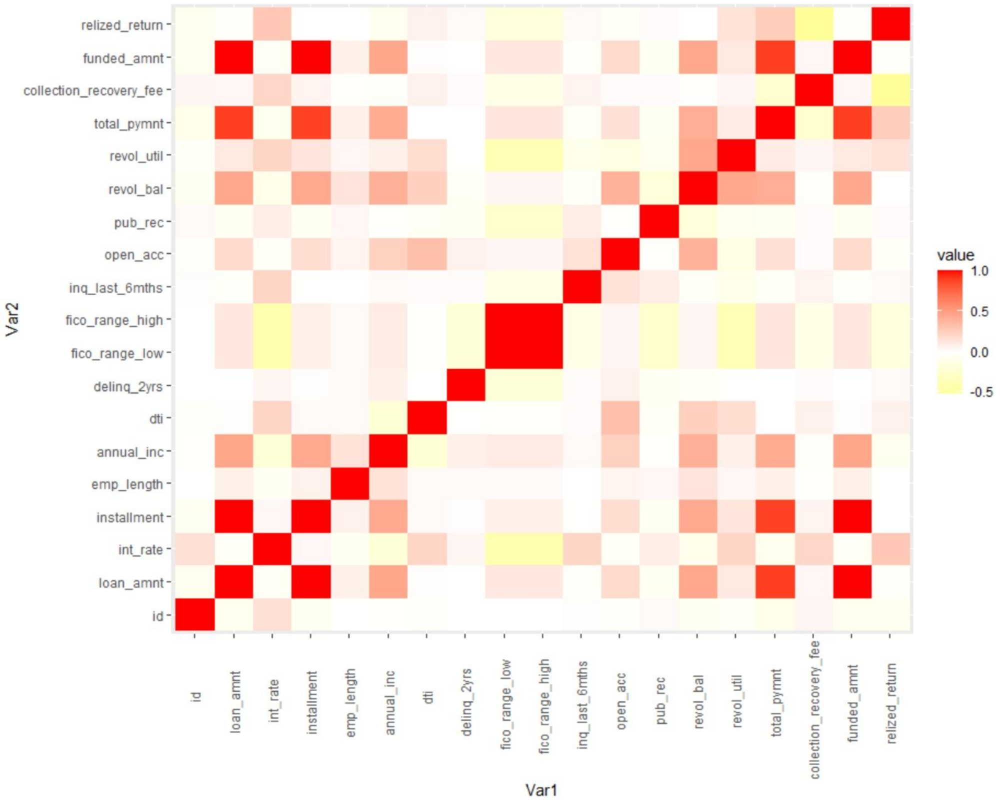


- 2.3 Distribution of loan term

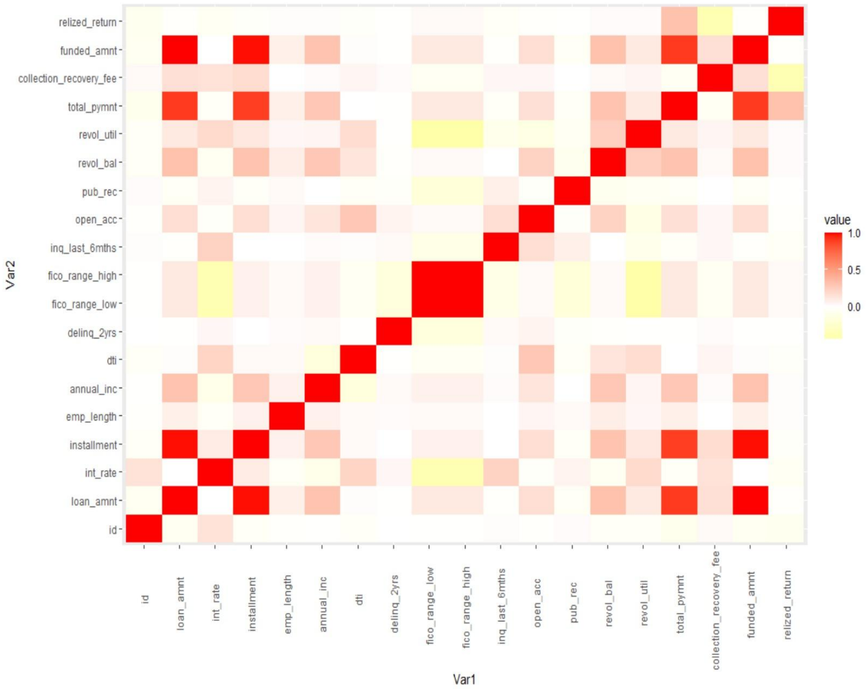


Correlation Graphs (2 of 6- only for example)

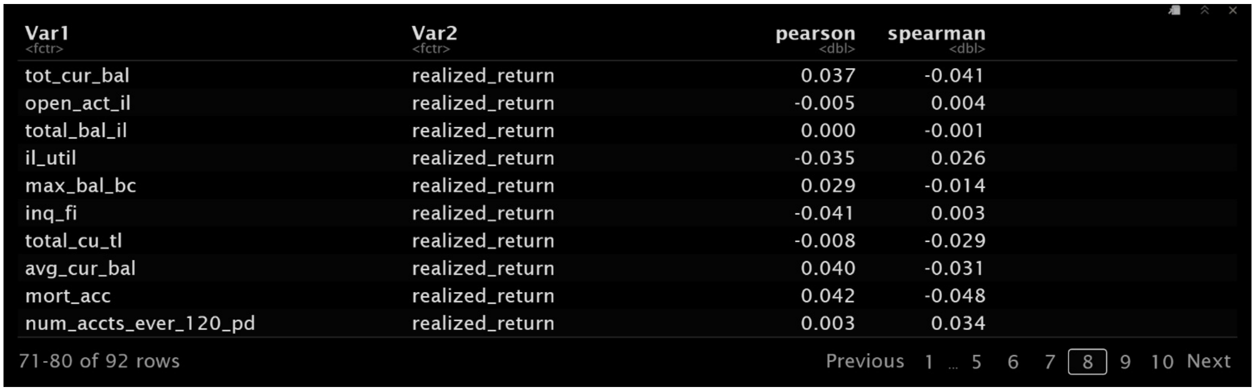
- 2.4 Spearman



- 2.5 Pearson



- 2.6 Example for the correlation between main variables to all variablels



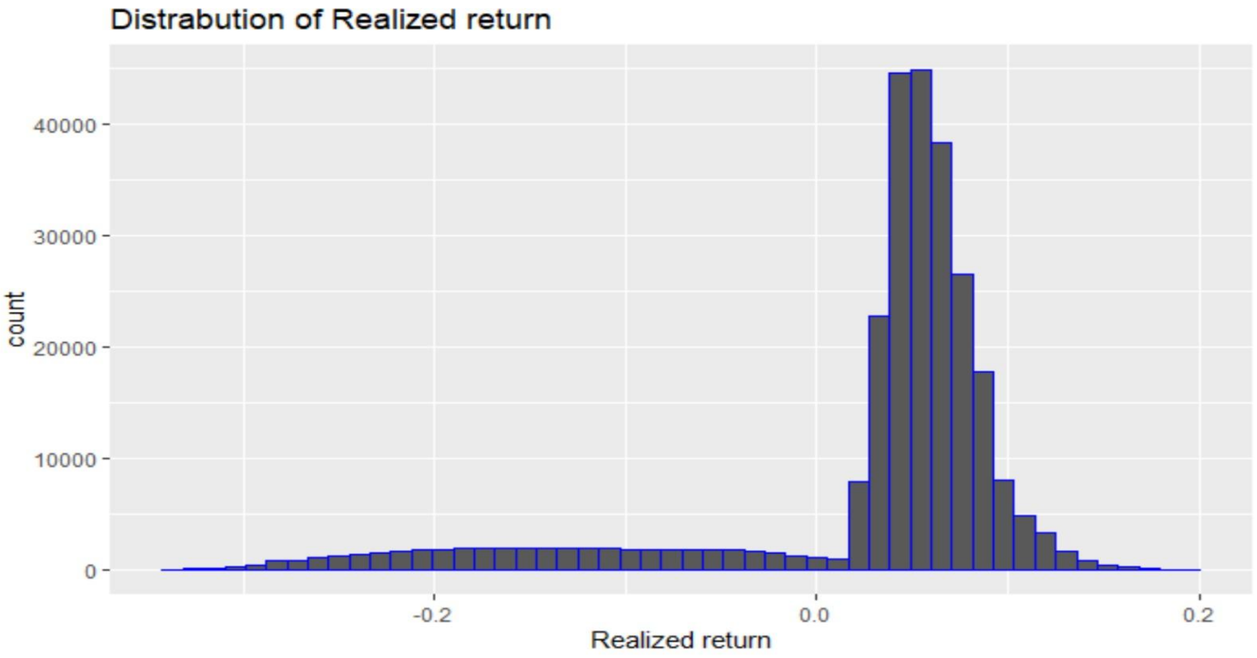
Var1 <fctr>	Var2 <fctr>	pearson <dbl>	spearman <dbl>
tot_cur_bal	realized_return	0.037	-0.041
open_act_il	realized_return	-0.005	0.004
total_bal_il	realized_return	0.000	-0.001
il_util	realized_return	-0.035	0.026
max_bal_bc	realized_return	0.029	-0.014
inq_fi	realized_return	-0.041	0.003
total_cu_tl	realized_return	-0.008	-0.029
avg_cur_bal	realized_return	0.040	-0.031
mort_acc	realized_return	0.042	-0.048
num_accts_ever_120_pd	realized_return	0.003	0.034

71-80 of 92 rows

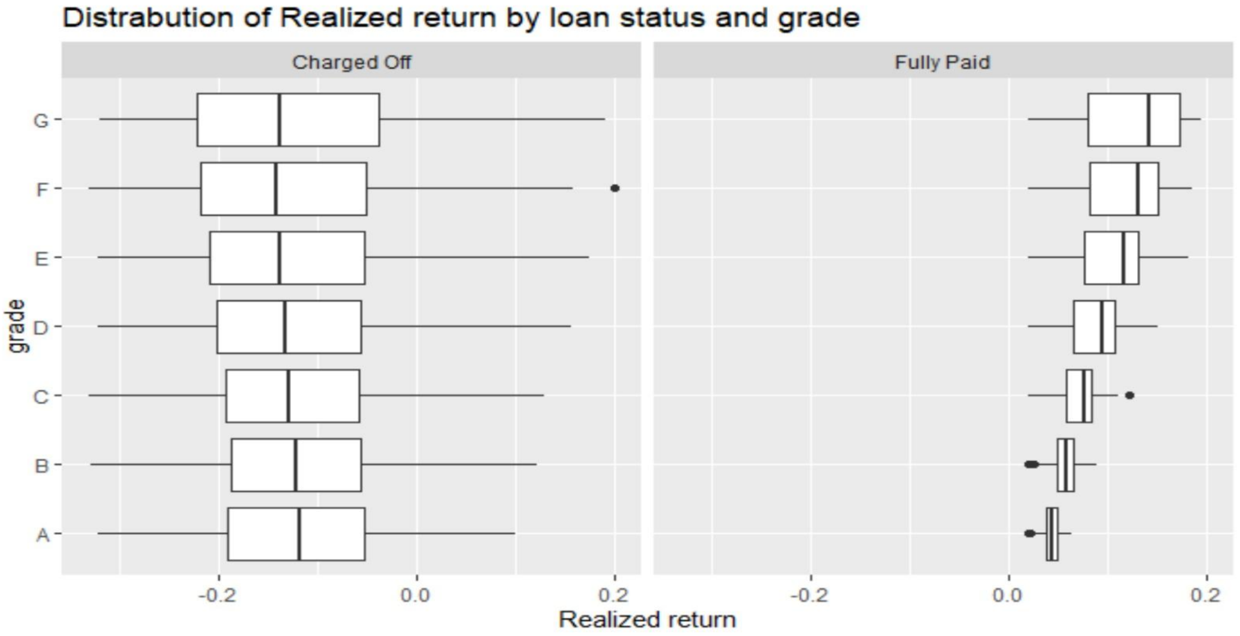
Previous 1 ... 5 6 7 8 9 10 Next

- Graphs of the realized return-

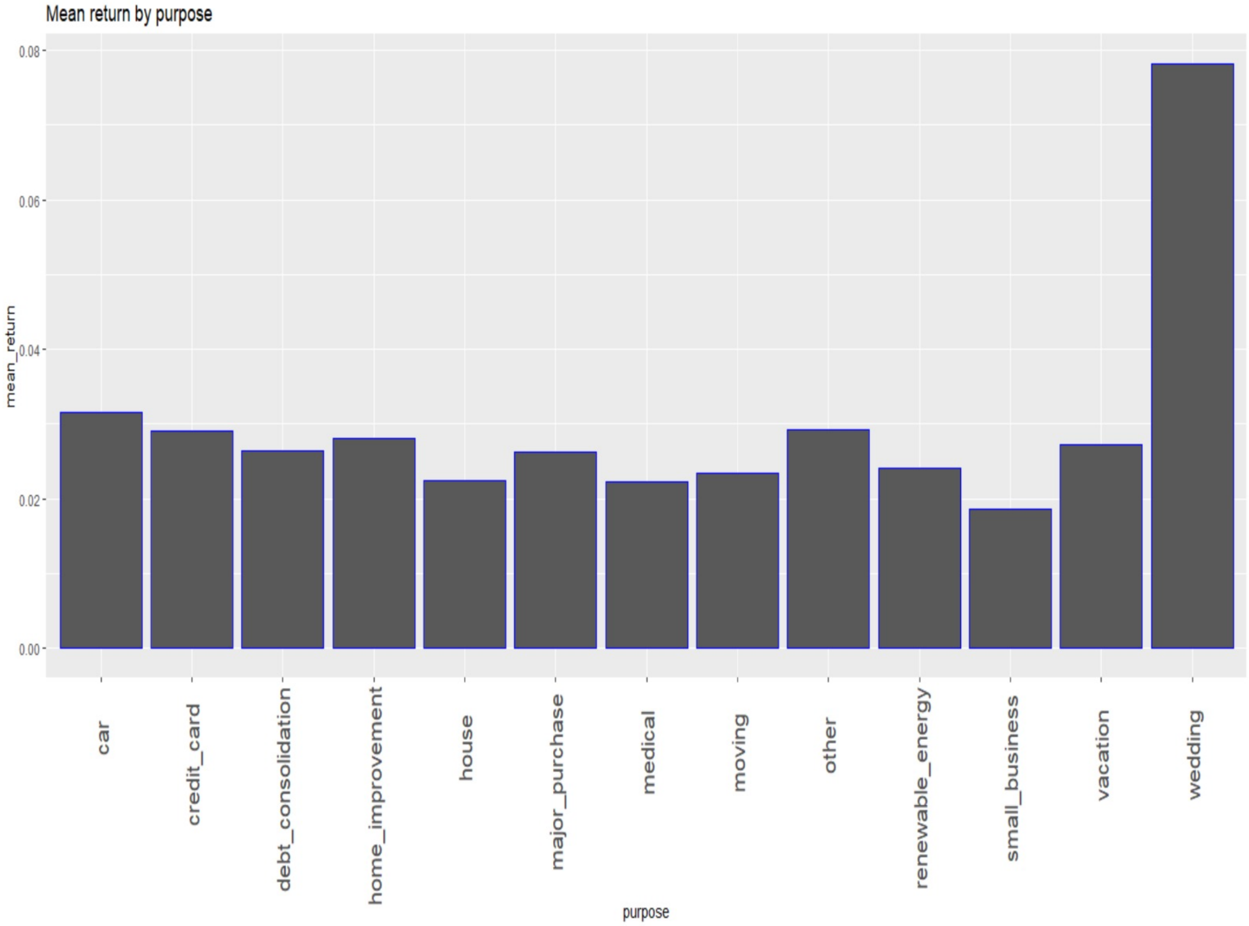
2.7



2.8



- 2.9 Mean Return by Purpose-


Mean return by purpose
mean_return
car
credit_card
debt_consolidation
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding
purpose