Machine Learning for Problem Solving
Case Study Phase 2: Data Cleaning, Preparation & Exploration

Zirui Zheng, ziruizhe
Keyu Zhang, keyuzhan
David Lee, dlee3

1. *Read the dataset that you downloaded in Phase I into Python. You will notice the data are provided in the form of many individual files, each spanning a certain time period. Combine the different files into a single data set.*

```
# Ingest the set of files we downloaded using the defined method "ingest_files"
files_cs = ingest_files(dir_cs) # dictioary of (filename, dataframe) as (key, value)
```

Combine the files

```
#print(list(files_cs.keys()))
data_cs = pd.concat(list(files_cs.values())) # combine "files_cs" into a pandas dataframe
data_cs = data_cs.reset_index(drop=True)
             # resent index with drop = True
```

See Jupyter Notebook report.

2. *Remove all instances (in our case, rows in the data table) representing loans that are still current (i.e., that are not in status Fully Paid, Charged-Off, or Default), and all loans that were issued before January 1, 2010.*

```
# Only include loans isssued since 2010
n_rows = len(final_data_3)

final_data_4 = final_data_3[final_data_3['issue_d'] >= datetime.date(2010,1,1)]

print("Removed " + str(n_rows - len(final_data_4)) + " rows")
```

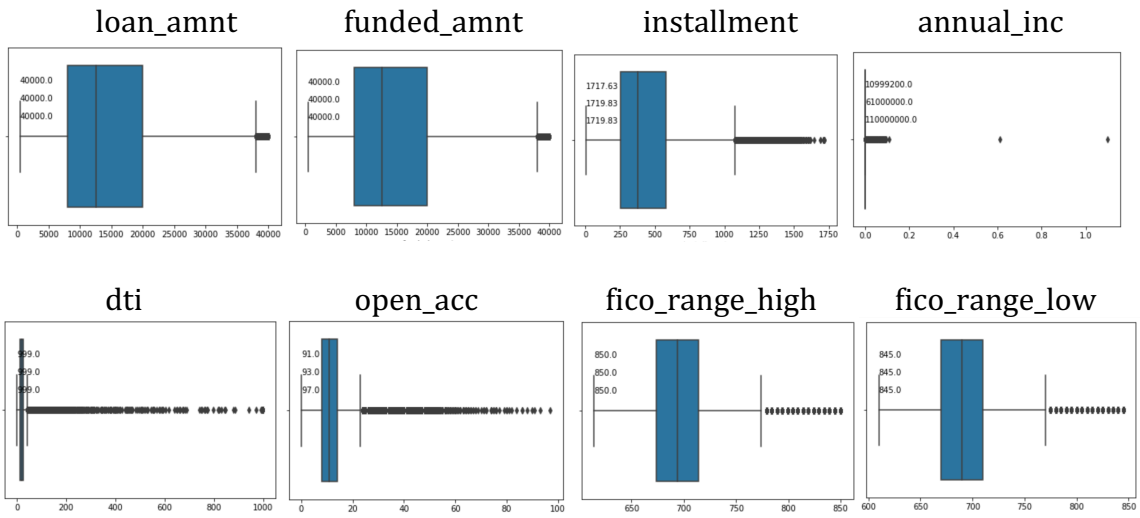
```
Removed 7340 rows
```

See Jupyter Notebook report.

3. *Visualize each of the features in the file. Are there any outliers? If yes, remove those instances.*

We've visualized continuous features by boxplots, categorical features by listing unique values, and date features by plotting the density of dates.

Here are some plots of continuous features:



Some data of categorical features:

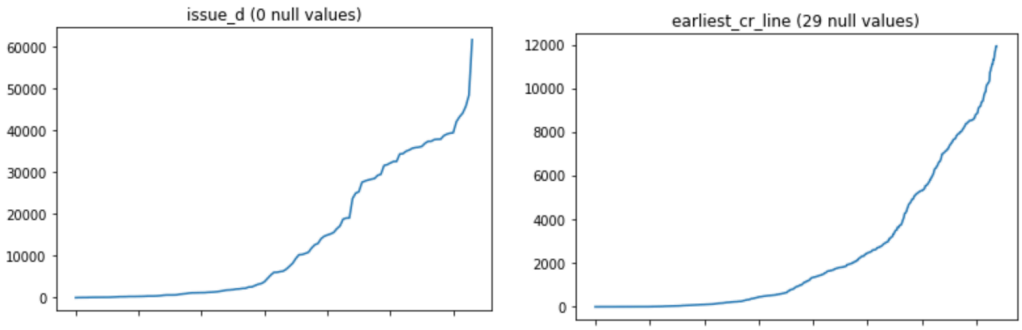
```
Field Name:  grade
Number of distinct values:  7
Distinct values and occurance:

C    521315
B    520527
A    296996
D    254390
E    116174
F     38424
G     11382
Name: grade, dtype: int64
```

```
Field Name:  loan_status
Number of distinct values:  9
Distinct values and occurance:

Current                                                792304
Fully Paid                                             736866
Charged Off                                            191939
Late (31-120 days)                                      20447
In Grace Period                                          8753
Late (16-30 days)                                        5758
Does not meet the credit policy. Status:Fully Paid       1988
Does not meet the credit policy. Status:Charged Off       761
Default                                                   392
Name: loan_status, dtype: int64
```

Some plots of date features:

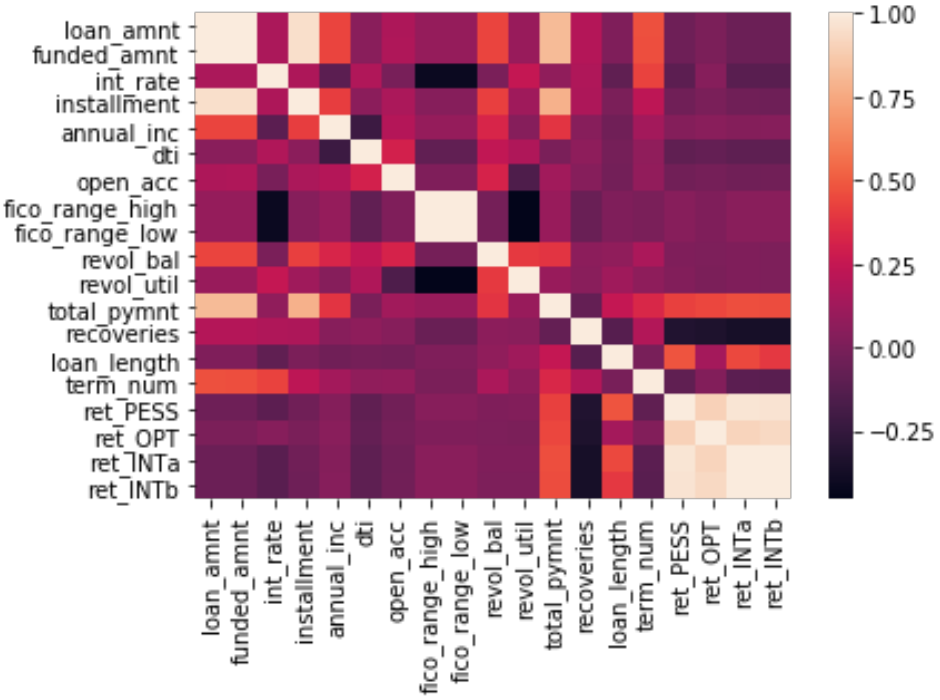


See Jupyter Notebook report for complete visualization of features.

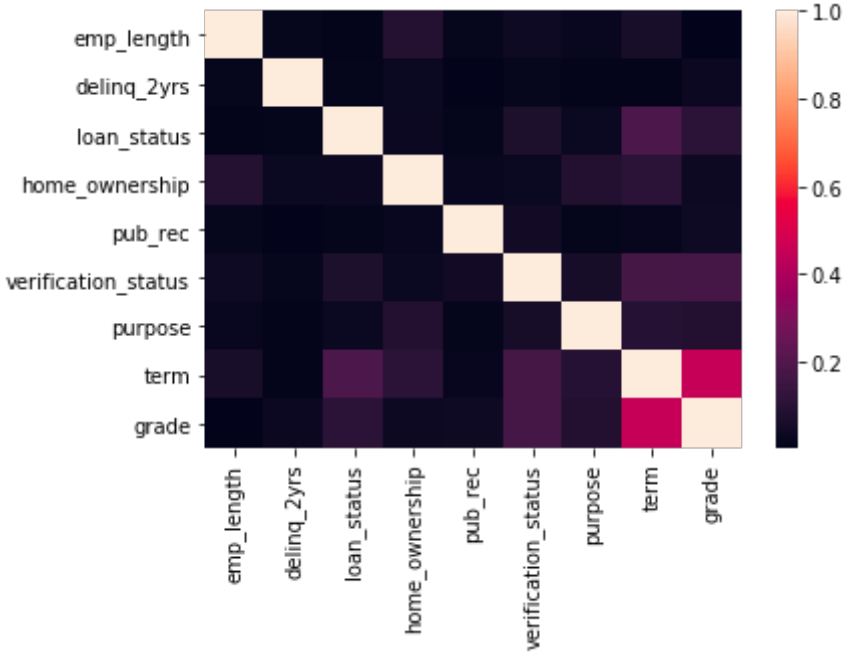
It is clear from the plots that there are outliers. We've removed outliers using the IQR method with a strict restriction of 1.5. Detailed method and visualization of cleaned data can be found in Jupyter Notebook report.

4. *Visually identify correlations between the features as well as the features and the loan status. Write down your observations.*

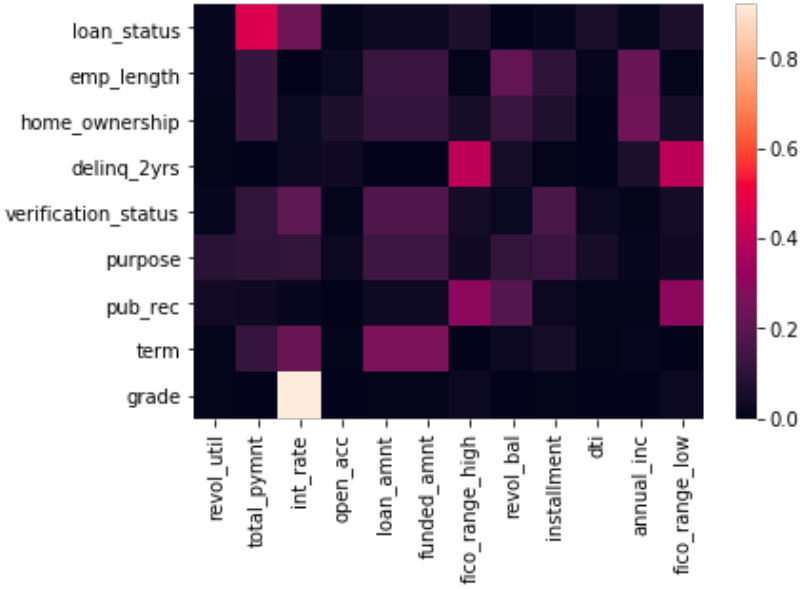
Correlation between numerical variables:



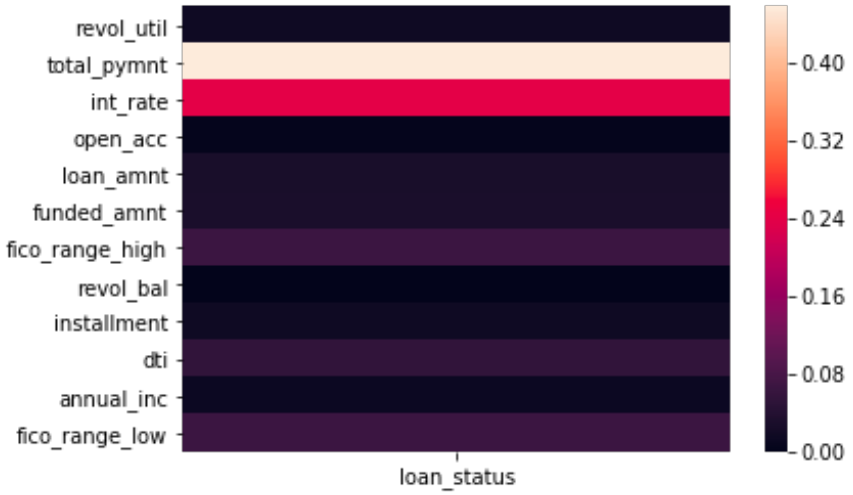
Correlation between categorical variables:



Correlation between categorical and numerical features:



Correlation between loan status and features:



See Jupyter Notebook report for more details.

5. *For the sake of this Study, restrict yourself to the following features: id, loan amnt, funded amnt, term, int rate, installment, grade, emp length, home ownership, annual inc, verification status, issue d, loan status, purpose, dti, delinq 2yrs, earliest cr line, open acc, pub rec, fico range high, fico range low, revol bal, revol util, total pymnt, last pymnt d, recoveries. Save the resulting data set in a Python "pickle*".

   See Jupyter Notebook report.

6. *Implement all three individual variables introduced in 6a.–6c. above (as given in Eq.s (2.1), (2.2), (2.3)) and add them as new variables to the dataset.*

   We used funded_amnt as "f" in the functions as this makes the most sense. See Jupyter Notebook report for method implementation.

7.

	perc_of_loans	perc_default	avg_int_rate	return_OPT	return_PESS	return_INTa	return_INTb
A	15.834194	7.486906	7.228165	3.673597	1.495293	3.028769	6.633429
B	29.013736	15.824236	10.868295	4.526283	1.212291	2.852341	6.441810
C	28.457783	26.143150	14.080998	4.564840	0.037382	2.011003	5.499790
D	15.609488	34.250551	17.558846	4.823840	-0.483460	1.525239	4.923266
E	7.629713	43.030598	20.784602	4.981762	-1.369943	0.671365	3.932414
F	2.746572	49.088053	24.511659	5.400009	-1.776046	0.177617	3.359229
G	0.708514	54.888363	27.166465	4.477794	-3.449531	-1.279702	1.738105

i. Percentage of loans for each grade:

A: 15.83% B: 29.01% C: 28.46% D: 15.61% E: 7.63%
F: 2.75% G: 0.71%

ii. Default rate for each grade:

A: 7.49% B: 15.82% C: 26.14% D: 35.25% E: 43.03%
F: 49.09% G: 54.89%

Those default rates are consistent with the assumption that the lower the grade, the greater the default rate. But still, the overall default possibility is surprisingly high, especially for the A and B level loan to have such a large default rate.

iii. Interest rate for each grade:

A: 7.22% B: 10.87% C: 14.09% D: 17.56%
E: 20.78% F: 24.51% G: 27.17%

The interest rates are consistent with the assumption that the lower the grade, the larger the interest rate. It's reasonable to find that these high interest rates are a good compensation for the high default possibility of those loans.

iv. Average return rates for each grade:

	OPTIMISTIC	PESSIMISTIC	M3(i=0.2%)	M3(i=0.5%)
A:	3.67%	1.50%	3.03%	6.63%
B:	4.53%	1.21%	2.85%	6.44%
C:	4.56%	0.04%	2.01%	5.50%
D:	4.82%	-0.48%	1.53%	4.92%
E:	4.98%	-1.37%	0.67%	3.93%
F:	5.40%	-1.78%	0.18%	3.36%
G:	4.48%	-3.45%	-1.28%	1.74%

v. No. It's normal to see the average return distributed like that. For the optimistic condition, the higher-grade loans are more likely to get paid back early, so the loose reinvestment assumption compensates the return rate a little bit to make the return close to the lower grade' returns. Under the other three strict reinvestment hypotheses, it's normal to see the return ordered along the grade.

From our point of view, we believe the M3 condition with a 2.4% reinvestment rate would be more realistic. Thus, for risk-averse investors like us, we would like to invest A-grade loans to ensure our return.