Carolina Carvalho Manhães Leite (leite2)
IE598 MLF F19
10/17/2019
Module 3 Homework (EDA)

Perform an exploratory data analysis (EDA) on the "High Yield Corporate Bond" dataset. Use the code listings presented in Bowles Chapter 2 to guide you.

The first analysis I performed were to count the number of rows and columns of the dataset (we have 2.721 rows and 37 columns). Following this step, I obtained the names of the variables:

CUSIP
Ticker
Issue Date
Maturity
1st Call Date
Moodys
S_and_P
Fitch
Bloomberg Composite Rating
Coupon
Issued Amount
Maturity Type
Coupon Type
Maturity At Issue months
Industry
LiquidityScore
Months in JNK
Months in HYG
Months in Both
IN_ETF
LIQ SCORE
n_trades
volume_trades
total_median_size
total_mean_size
n_days_trade
days_diff_max
percent_intra_dealer
percent_uncapped
bond_type
Client_Trade_Percentage
weekly_mean_volume
weekly_median_volume
weekly_max_volume

weekly_min_volume
weekly_mean_ntrades
weekly_median_ntrades

From those names we can already predict some behaviors. Let's see if these hypothesis hold:

- Some features with high correlation (all of the "weekly_volume" variables amongst each other, with the same happening for "total_size" and "LIQ SCORE" vs "LiquidityScore";
- Even though some features are represented by integers, they behave like categorical variables (bond_type)

I also noticed we have a few variables with ratings attributed by different agencies. Imagining that, in the future, this dataset may be the source of information for some model predicting those ratings, I aggregated the Rating Bloomberg (which is some sort of combination of the others) in fewer classes: AAA/AA/A, BBB/BB/B, CCC/CC/C, DDD/DD/D and Others. Calculating their frequencies:

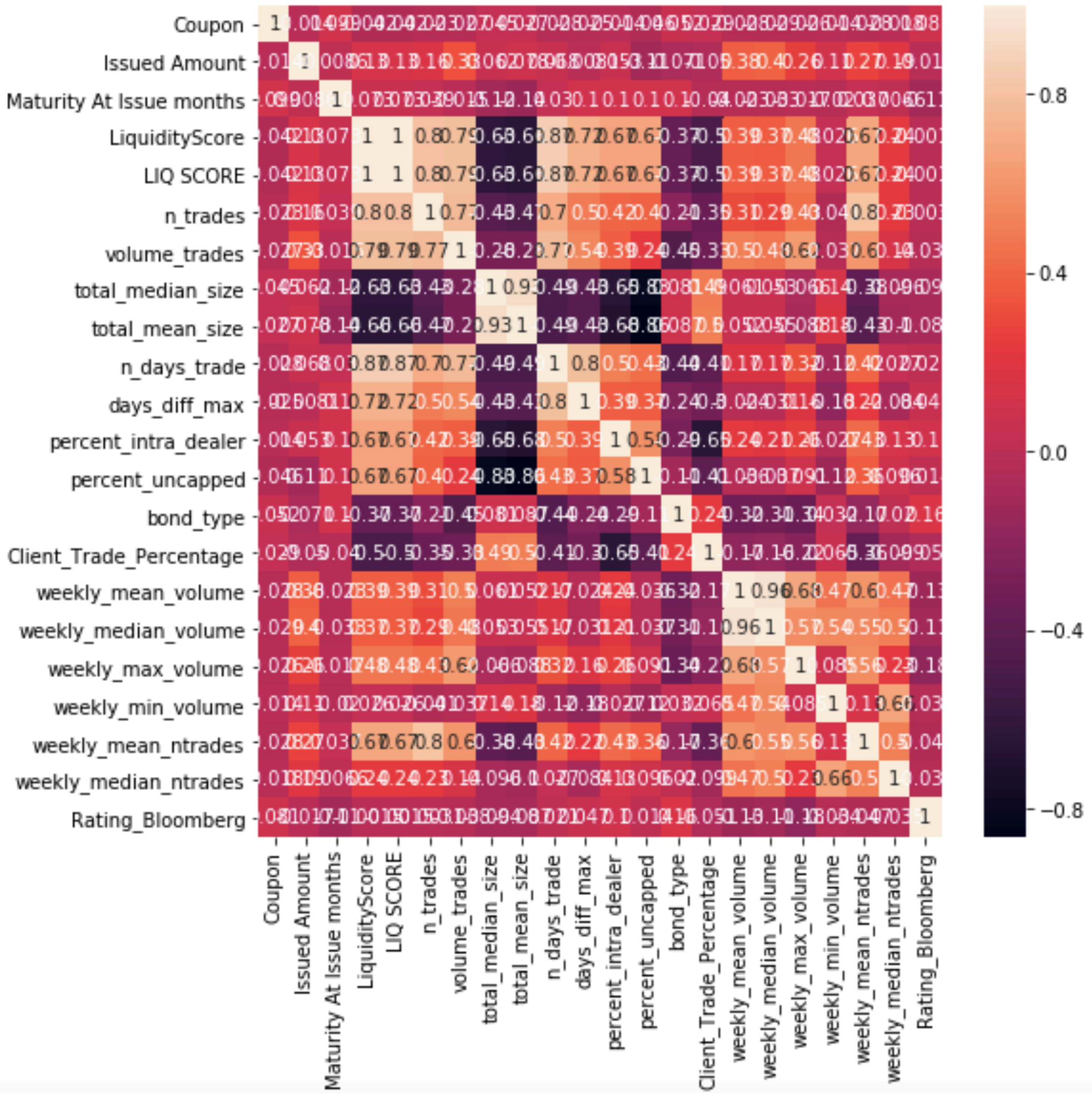
Rating (Aggregated)	Frequency (# obs)	Frequency (%)
AAA/AA/A (1 in the code)	139	5.1%
BBB/BB/B (2 in the code)	1230	45.2%
CCC/CC/C (3 in the code)	171	6.3%
DDD/DD/D (4 in the code)	4	0.1%
Others (5 in the code)	1177	43.3%

If we were to actually fit a model we'd have to better analyze the number of missing values in the target variable; "others" ("Nan" and "NR") makes up for almost half of the sample. For our purposes here, I'll let them stay in our dataset.

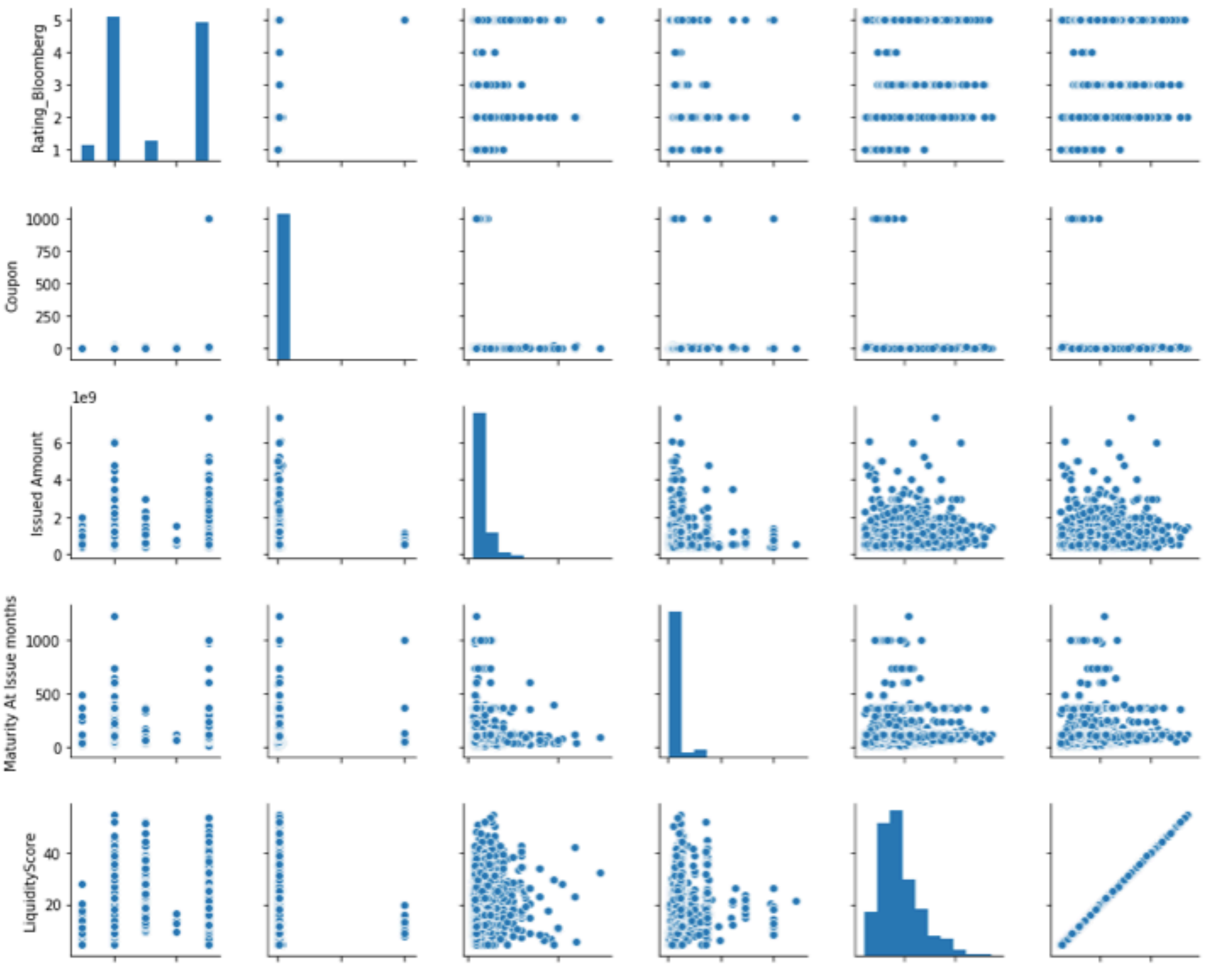
Following this step, I performed some descriptive analysis of the variables in our dataset; for the continuous ones, the method "describe" calculated means, medians, quartiles, max and min and, for the categorical ones, the number of unique values and the domain with the higher frequency. Those tables can help us see if the data have very different ranges (which could lead to some future standardization process if we're thinking about using PCA, Logistic Regression or other technique better handled with normalized data). We can also look for outliers, given the maximum and minimum values for each attribute. Since it's an exhaustive search I won't display all of the results here in the report (they're in my code, though), but some conclusions we can draw from it are:

- The variable "CUSIP" has 2721 unique values (exactly the same number of rows of the whole dataset). It means this is probably an index and should not be considered while fitting a model (variables like these would dangerously make up for "perfect features" in terms of discriminating classes… after all they're different for each one of the observations);
- Some variables have special missing values ("Maturity" is supposed to be a date, but it has "Nan Field Not Applicable" as one of its domains). This illustrates the fact that not all missing values are actually "missing": they can be represented by special domains.

We can also compute the correlation matrix and display it as a heatmap:

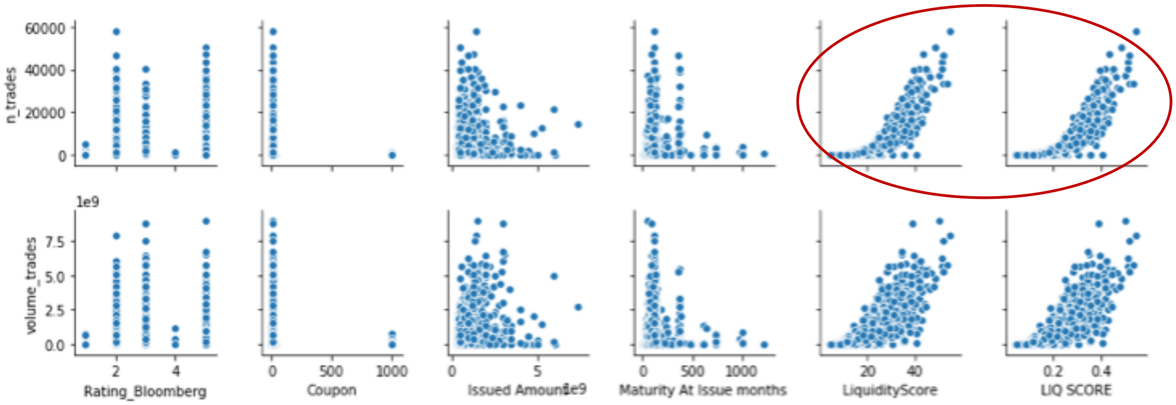


Very light colors and very dark ones represent high positive and negative linear correlations, respectively. As expected, the main diagonal has the correlations of the variables with themselves (i.e., 1); we also can see that the variables of the same "type/idea" (like "total_size") have higher correlations. Another thing we can infer from the plot is that, actually, none of the features have a very high correlation with the target I created ("Rating_Bloomberg"); we can see that the last row is predominantly pink (the middle of the scale). It doesn't necessarily mean that there's no relation at all between the feaures and the rating, it just means there's no linear correlation. A tool to help us see if there're non-linear relationships is the scatterplot matrix. Again, I won't display all of the combinations here (they're already in the code), but just from a sample we can infer some conclusions:



- “Rating_Bloomberg” is actually splitted in the following way: “half B’s and half missing values”. The missing values are a real concern here;
- Coupon seems to have an outlier around 1000: it’s the point standing out in the plot;
- Other variables are presented just like some cloud of data…
- “LiquidityScore” and “LIQ SCORE” have almost a perfect linear relationship, as expected.

Another interesting part of the scatterplot:



The plots circled in red may indicate that those pairs of variables have a “well-behaved” relationship. We could try some transformations to see which one fits the data the best (exponential could be an option).

Part 2: Appendix

https://github.com/leiteccml/Carolina_2019/tree/master/IE598_F19_HW3