Business Problem

- Given a dataset filled with demographic, meta and financial data regarding a customer, come up with a data solution that will aid the company to **maximize its profits and minimize its losses.**

Outline for How to Proceed

- **Given I were the manager of this Data Science project at **Company** tasked with this problem:**
  - Communicate with the data engineers who are in charge of the ETL to figure out the schema of the data.
  - Communicate with business experts and analysts on the team to figure out the meaning behind each individual field. Work together to figure out best **features that we are looking to optimize** (or minimize) before building a model.
  - Communicate with business experts to figure out the **best metrics in determining success** / evaluating our model.
  - Perform Data preprocessing and feature engineering.
  - Train and test data with a chosen algorithm, repeat the above steps again to iterate on the process until it is fine tuned.

Exploratory Data Analysis

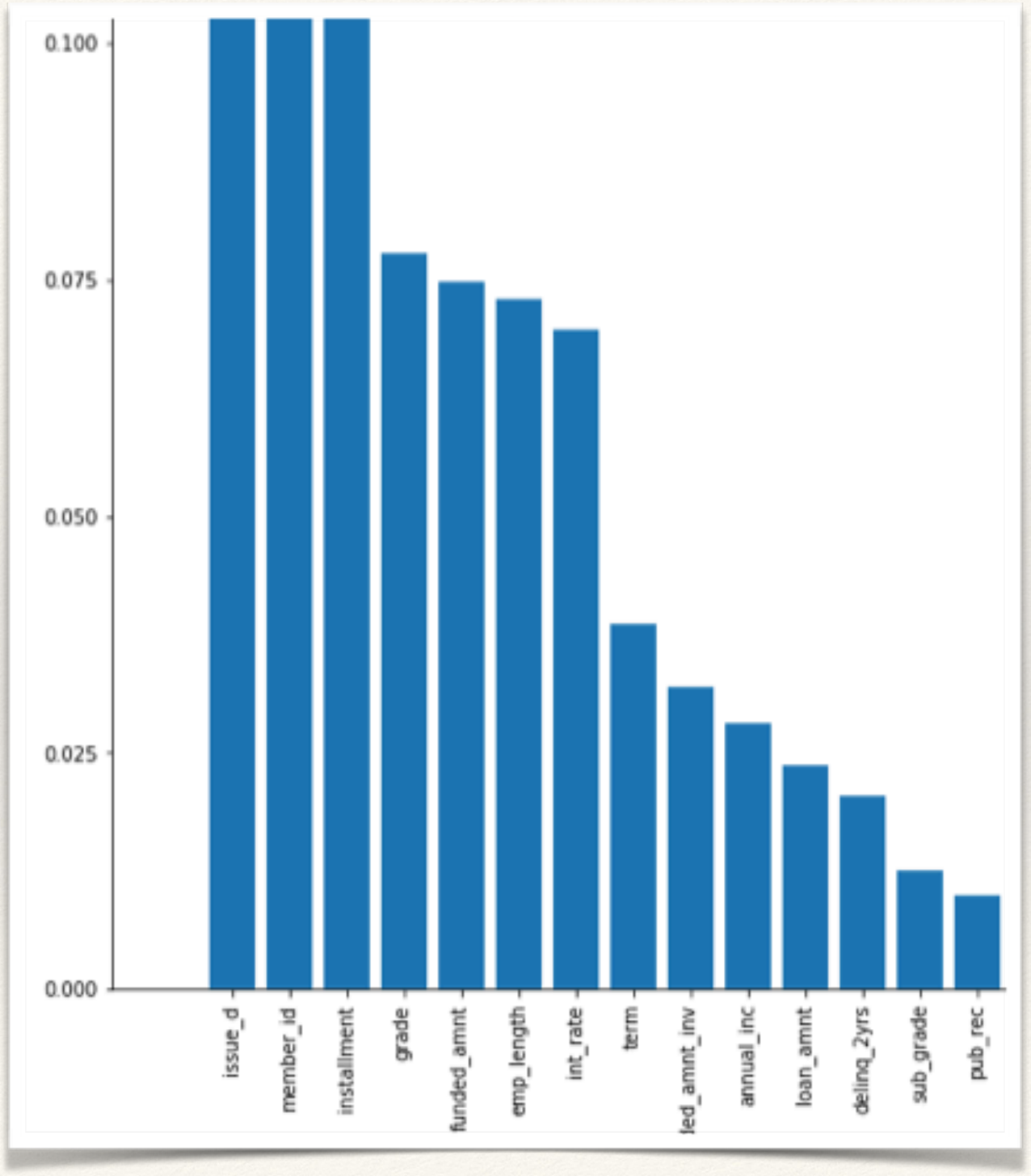
```
df = pd.read_csv('lc_loan.csv')
df.head()
```

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	total_bal_il	il_util	open_rv_12m	open_r
0	1077501	1296599	5000	5000	4975.0	36 months	10.65	162.87	B	B2	...	NaN	NaN	NaN	
1	1077430	1314167	2500	2500	2500.0	60 months	15.27	59.83	C	C4	...	NaN	NaN	NaN	
2	1077175	1313524	2400	2400	2400.0	36 months	15.96	84.33	C	C5	...	NaN	NaN	NaN	
3	1076863	1277178	10000	10000	10000.0	36 months	13.49	339.31	C	C1	...	NaN	NaN	NaN	

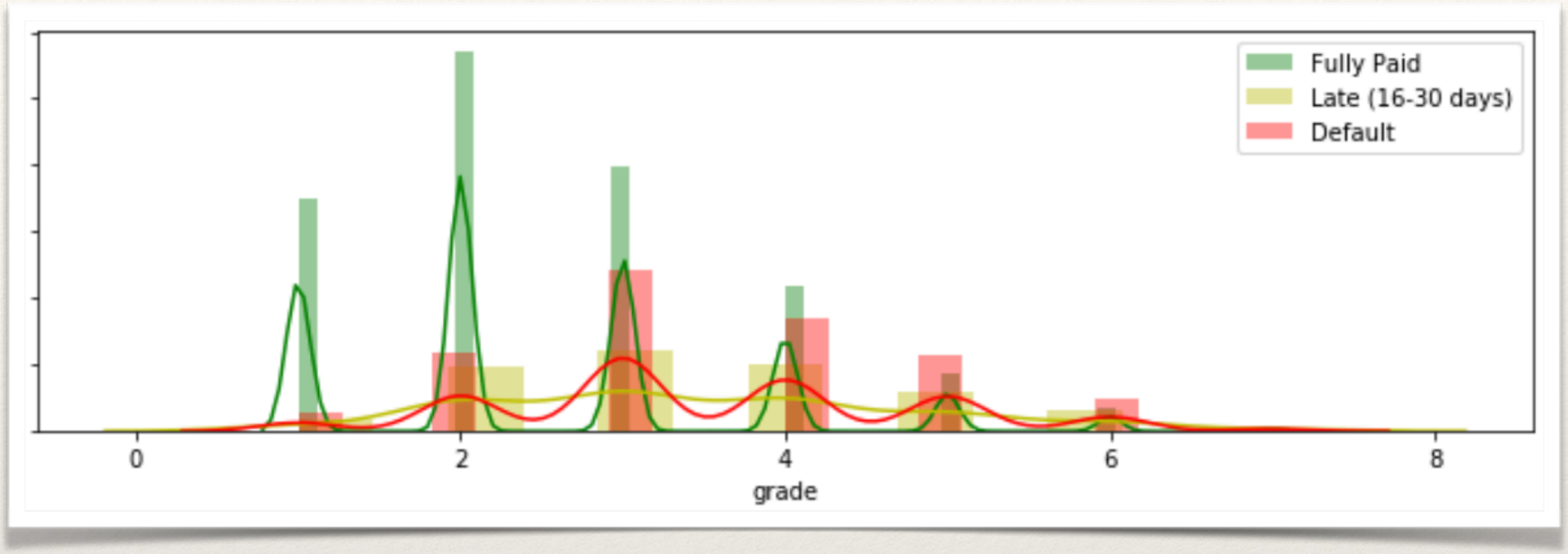
- Check for **null values** and fill or delete them.
  - Fill null values with statistical data such as mean, median, mode, max, min, or 0, depending on the business meaning of the data.
- Deal with **categorical fields**. Only numerical values can be entered into an algorithm, so we must transform categories to numbers in a statistically meaningful way.
- **Correlation** / Variance Inflation Analysis

Feature Extraction

- Features have been extracted from our dataset in several ways:
  - First, we removed features that would be unknown at the start of a customer's loan (such as number of late payments or total interest paid)
- Second, we removed features with high multicollinearity.
- Third, if we are looking at **CART** models, they do feature extraction naturally.

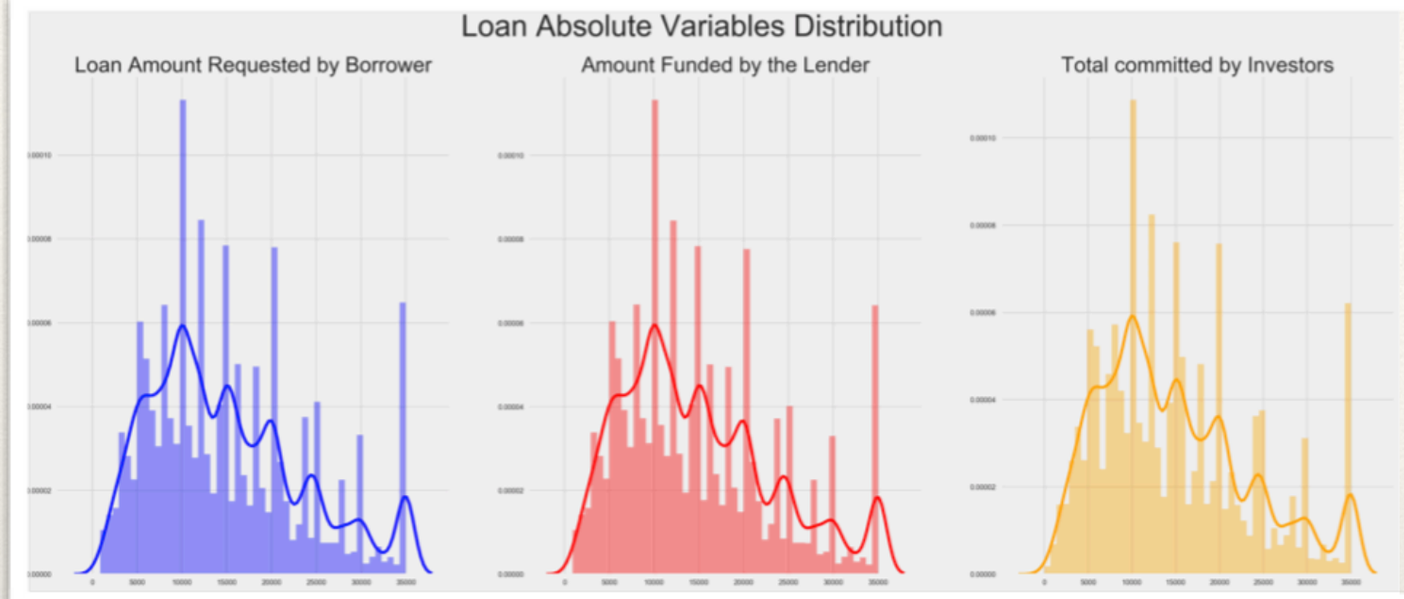


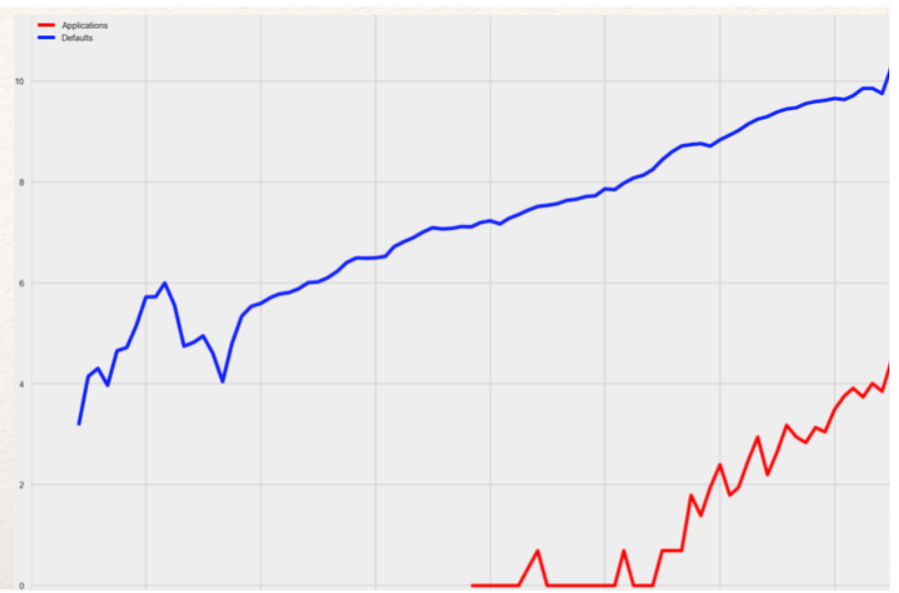
Visualization



- **Examining graphs of relevant features categorized by their loan status, we can see that there are clear feature distinctions between the label classes.**

Important Visualizations:





- **Loan Absolute Variables Distribution: Amt Requested, Funded by lender, committed by investor**
- **Total Applications Volume vs Defaulted Volume**
- **Loan Purpose**
- **Average Interest Rates**
- **Delinquencies**

Model Building

- **Back to the business problem at hand**: given our current knowledge of the data we have, how can we figure out a model that will ultimately maximize profitability and minimize loss potential.
  - **Option 1:** No intrinsic labels to our data, can treat it as an **Unsupervised Learning Problem** - cluster customers into 3 categories of “Safe”, “Risky”, and “Unsure”.
  - **Option 2:** Pick some feature / combination of features to predict.
    - ***loan_status*:**
      - 'Fully Paid’
      - 'Charged Off’
      - 'Current', 'Default'
      - 'Late (31-120 days)’
      - 'In Grace Period’
      - 'Late (16-30 days)',
      - 'Does not meet the credit policy. Status:Fully Paid',
      - 'Does not meet the credit policy. Status:Charged Off'

Model Building

- More Sophisticated Method: working together with business experts, come up with a selection of features to predict on and build a regression or classification model around this engineered feature.
  - Could be categorical (grouping members into categories to define which loans to accept) or regression (deciding the amount of money to loan a particular person)

Defining Accuracy Metrics

- Unsupervised Learning: Silhouette method
- Supervised Learning:
  - Classification: Decide between F1, accuracy, ROC-AUC depending how important false positives/ negatives are.
  - Regression: Decide between R^2, RMSE, MAE, depending on how much we care to predict/ take into account outliers.