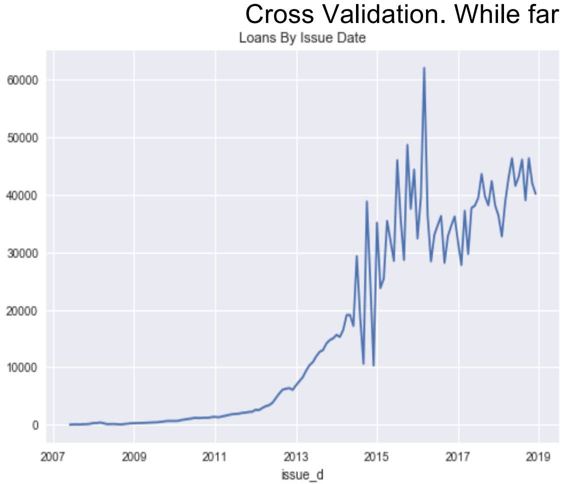
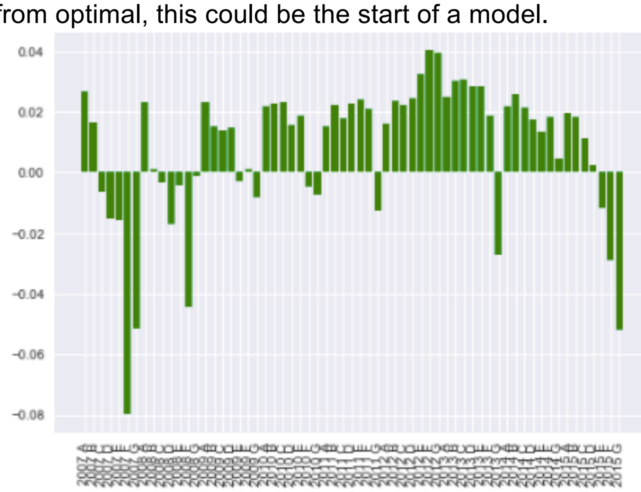
Lending Club – Data Analysis
Net/Net:

- **Data Exploration:** The primary feature that needs adjustment is income, which is handled by removing nulls as well as the 1st and 99th percentiles. The variable must then be logged as to make it normally distributed. Small changes are also made to other variables (i.e. enforce DTI > 0).
- **Business Analysis**
    1. What percent of loans has been fully paid? 86.05%
    2. When bucketed by year of origination and grade which cohort has the highest rate of defaults? 2015-G
    3. When bucketed by year of origination and grade, what annualized rate of return have the loans generated on average? Full Data: 1.90%, Average of Individual Year of Origination X Grade Cohorts: 0.79%
- **Model**
    - *Pre-Model Evaluation*
        - ***The grade assigned by Lending Club has the highest correlation with the default indicator. EDA showed the lending club grade did a good job assigning grades to relative default rates.***
        - ***Grade and interest rate are closely related due to the underlying factors that are likely modeling both features.***
        - Funding is largely related to income levels
        - The credit revolver is related to both income levels and the dti ratio
        - There is not a strong relationship between income and dti
    - *Logistic Regression Formulation*
        - Because defaulted loans are under-represented in the dataset leverage SMOTE (Synthetic Minority Oversampling Technique)
        - Check the relative ranking of features using RFE (Recursive Feature Elimination)
        - Following testing, build the logistic regression with the following 4 variables: ['int_rate','grade_cat','log_inc','dti']
        - The variables' signs have practical interpretations
            - Worse grade leads to a higher likelihood of default
            - Higher interest rates lead to a higher likelihood of default
            - Higher income leads to a lower likelihood of default
            - Higher Debt to Income ratios lead to a higher likelihood of default
    - *Model Performance*
        - The model is of low quality (r-squared: 0.064) classifying 63% of cases correctly (note: this is much improved compared to without using SMOTE). The robustness of results is confirmed with a K-Fold Cross Validation. While far from optimal, this could be the start of a model.





Origination_Year	defaulted	loans	default_rate
2015	41532	277908	14.94
2014	21926	159947	13.71
2013	12186	99225	12.28
2012	5820	42843	13.58
2011	1458	13796	10.57
2010	948	8889	10.66
2009	692	5114	13.53
2008	480	2335	20.56
2007	133	534	24.91

Origination_Year	annualized_rate_of_return
2007	-0.010923
2008	-0.003214
2009	0.014982
2010	0.021042
2011	0.018909
2012	0.021747
2013	0.028977
2014	0.022106
2015	0.013473

grade	defaulted	loans	default_rate
G	248	627	39.55
F	1408	4234	33.25
E	6033	20990	28.74
D	17542	74799	23.45
C	28685	159824	17.95
B	23517	209313	11.24
A	7742	140804	5.50

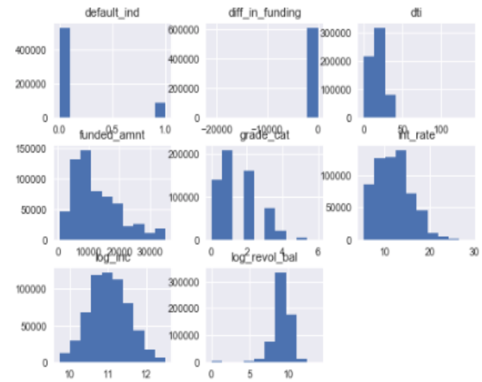
grade	annualized_rate_of_return
A	0.020342
B	0.022933
C	0.017801
D	0.013222
E	0.005222
F	0.000699
G	-0.020074

Part 1: Data Exploration and Evaluation

The data represents a point in time snapshot of individual loans (rows)…initial exploration of the data reveals:

1) There are nulls in the annual income and dti fields
2) There are outliers in income (both low (0) and high) as well as dti (values < 0) and revol_bal
3) Fields such as loan_status must be unified to a common outcome description or default indicator

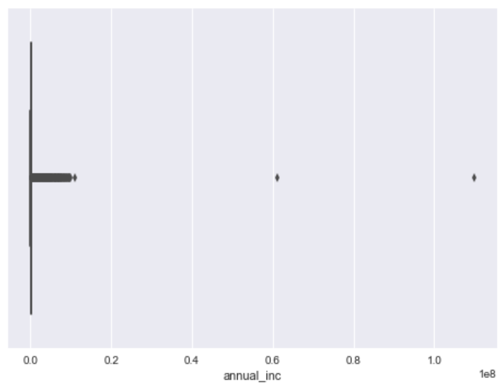
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
loan_amnt	2.26067e+06	NaN	NaN	NaN	15046.9	9190.25	500	8000	12900	20000	40000
funded_amnt	2.26067e+06	NaN	NaN	NaN	15041.7	9188.41	500	8000	12875	20000	40000
term	2260668	2	36 months	1609754	NaN	NaN	NaN	NaN	NaN	NaN	NaN
int_rate	2.26067e+06	NaN	NaN	NaN	13.0929	4.83211	5.31	9.49	12.62	15.99	30.99
grade	2260668	7	B	663557	NaN	NaN	NaN	NaN	NaN	NaN	NaN
annual_inc	2.26066e+06	NaN	NaN	NaN	77992.4	112696	0	46000	65000	93000	1.1e+08
issue_d	2260668	139	Mar-2016	61992	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dti	2.25896e+06	NaN	NaN	NaN	18.8242	14.1833	-1	11.89	17.84	24.49	999
revol_bal	2.26067e+06	NaN	NaN	NaN	16658.5	22948.3	0	5950	11324	20246	2.90484e+06
total_pymnt	2.26067e+06	NaN	NaN	NaN	11824	9889.6	0	4272.58	9060.87	16708	63296.9
loan_status	2260668	9	Fully Paid	1041952	NaN	NaN	NaN	NaN	NaN	NaN	NaN



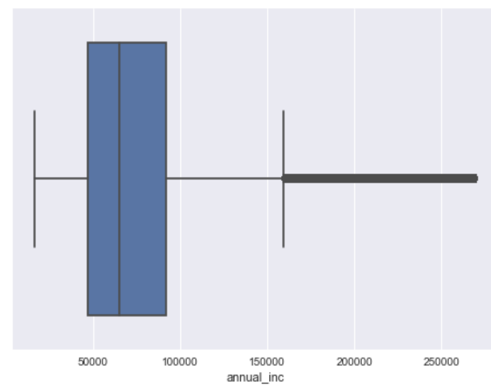

```
Total Rows: 2260668
```

Annual Income: 1) remove outliers on the low end (i.e. 0 – unreported) and high-end (i.e. fat finger error or intentional over statement) by using data between the 1st and 99th percentiles 2) Log the variable to create a normal distribution

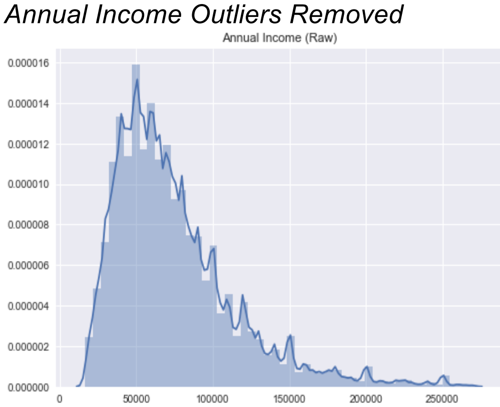
Raw Annual Income



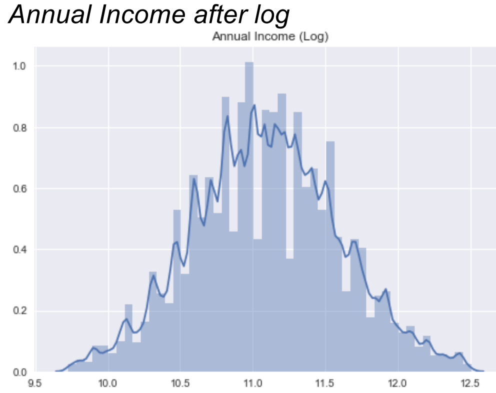
Annual Income Outliers Removed



Annual Income Outliers Removed

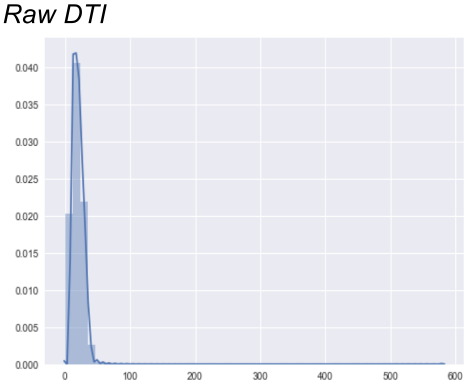


Annual Income after log

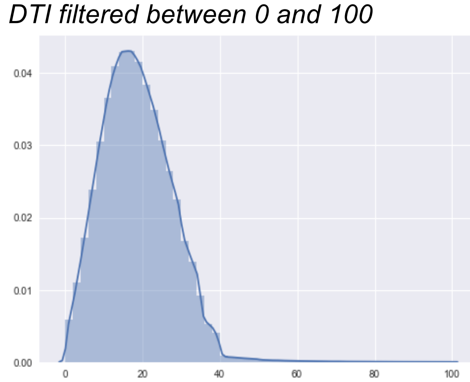


DTI: 1) enforce DTI > 0, 2) aware of DTIs > 100 but do not enforce a restriction (high DTI possible, though unlikely)

Raw DTI



DTI filtered between 0 and 100

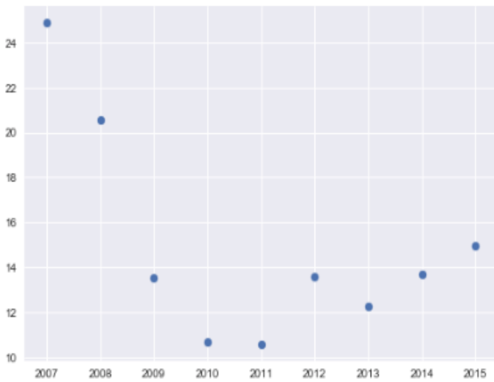


Part 2: Business Analysis

While marked as fully paid or charged off, the 2016-2018 loan cohorts likely do not have full 36-month term data as evidenced by their origination date and returns below reasonable expectations. Further unify the loan status into a simpler hierarchy; note: there is a small subset of loans that are still in process (i.e. late).

1) *What percent of loans has been fully paid? 86.05%*
2) *When bucketed by year of origination and grade which cohort has the highest rate of defaults? 2015-G*

Performance by Origination Year



Origination_Year	defaulted	loans	default_rate
2015	41532	277908	14.94
2014	21926	159947	13.71
2013	12186	99225	12.28
2012	5820	42843	13.58
2011	1458	13796	10.57
2010	948	8889	10.66
2009	692	5114	13.53
2008	480	2335	20.56
2007	133	534	24.91

Performance by Grade

grade	defaulted	loans	default_rate
G	248	627	39.55
F	1408	4234	33.25
E	6033	20990	28.74
D	17542	74799	23.45
C	28685	159824	17.95
B	23517	209313	11.24
A	7742	140804	5.50

Year of Origination X Grade

Origination_Year	grade	defaulted	loans	default_rate
2015	G	107	221	48.42
2007	F	24	51	47.06
2007	G	13	29	44.83
2015	F	552	1292	42.72
2009	G	18	49	36.73

3) *When bucketed by year of origination and grade, what annualized rate of return have the loans generated on average? Full Data: 1.90%, Average of Individual Year of Origination X Grade Cohorts: 0.79%*

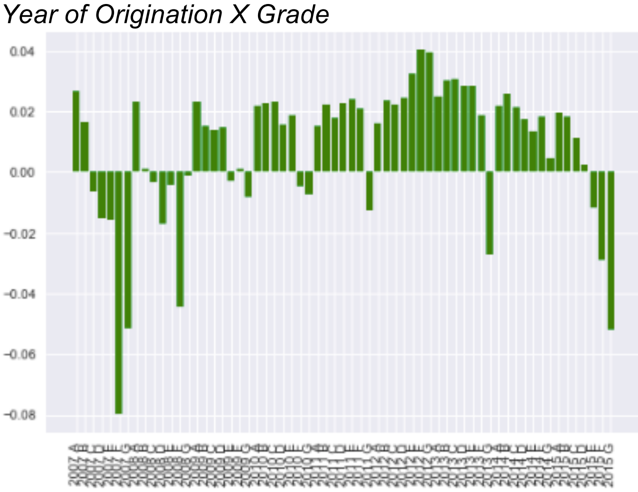
Year of Origination

Origination_Year	annualized_rate_of_return
2007	-0.010923
2008	-0.003214
2009	0.014982
2010	0.021042
2011	0.018909
2012	0.021747
2013	0.028977
2014	0.022106
2015	0.013473

Grade

grade	annualized_rate_of_return
A	0.020342
B	0.022933
C	0.017801
D	0.013222
E	0.005222
F	0.000699
G	-0.020074

Year of Origination X Grade



Part 3: Modeling (Building a Logistic Regression)

Note: For the model further filter to loans with definite outcomes i.e. either Fully Paid or Charged Off

Correlation Matrix

	funded_amnt	diff_in_funding	int_rate	grade_cat	log_inc	dti	log_revol_bal	default_ind
funded_amnt	1.000000	0.007680	-0.084115	-0.083594	0.497529	0.011973	0.384587	-0.015184
diff_in_funding	0.007680	1.000000	0.015430	0.006199	-0.010732	0.016266	-0.002447	0.000158
int_rate	-0.084115	0.015430	1.000000	0.939436	-0.214134	0.138035	-0.096716	0.188700
grade_cat	-0.083594	0.006199	0.939436	1.000000	-0.204665	0.145428	-0.105439	0.193332
log_inc	0.497529	-0.010732	-0.214134	-0.204665	1.000000	-0.216139	0.317277	-0.087823
dti	0.011973	0.016266	0.138035	0.145428	-0.216139	1.000000	0.227523	0.080762
log_revol_bal	0.384587	-0.002447	-0.096716	-0.105439	0.317277	0.227523	1.000000	-0.029943
default_ind	-0.015184	0.000158	0.188700	0.193332	-0.087823	0.080762	-0.029943	1.000000

- ***The grade assigned by Lending Club has the highest correlation with the default indicator***
- ***Grade and interest rate are closely related due to underlying factors that are likely modeling both features***
- Funding is largely related to income levels
- The credit revolver is related to both income levels and the dti ratio
- There is not a strong relationship between income and dti

Logistic Regression Formulation

- Because defaulted loans are under-represented leverage SMOTE (Synthetic Minority Oversampling Technique)
- Check the relative ranking of features using RFE (Recursive Feature Elimination)
- Following testing, build the logistic regression with the following 4 variables: ['int_rate','grade_cat','log_inc','dti']

```
Optimization terminated successfully.
         Current function value: 0.648831
         Iterations 5
                          Results: Logit
==================================================================
Model:              Logit            No. Iterations:   5.0000
Dependent Variable: default_ind      Pseudo R-squared: 0.064
Date:               2019-11-17 10:16 AIC:              955061.5852
No. Observations:   735980           BIC:              955107.6210
Df Model:           3                Log-Likelihood:   -4.7753e+05
Df Residuals:       735976           LL-Null:          -5.1014e+05
Converged:          1.0000           Scale:            1.0000
-------------------------------------------------------------------
             Coef.   Std.Err.      z      P>|z|   [0.025   0.975]
-------------------------------------------------------------------
int_rate     0.0834   0.0019    44.8275  0.0000   0.0798   0.0871
grade_cat    0.2244   0.0064    34.9122  0.0000   0.2118   0.2370
log_inc     -0.1612   0.0014  -117.6452  0.0000  -0.1639  -0.1585
dti          0.0186   0.0003    63.3591  0.0000   0.0181   0.0192
==================================================================
```

The variables' signs have practical interpretations

- Worse grade leads to a higher likelihood of default
- Higher interest rates lead to a higher likelihood of default
- Higher income leads to a lower likelihood of default
- Higher Debt to Income ratios lead to a higher likelihood of default

Model Performance

The model is of low quality (r-squared: 0.064) classifying 63% of cases correctly (note: this is much improved compared to without using SMOTE). The robustness of results is confirmed with a K-Fold Cross Validation. While far from optimal, this could be the start of a model.

```
              precision    recall  f1-score   support

           0       0.62      0.63      0.63    109989
           1       0.63      0.62      0.63    110805

    accuracy                           0.63    220794
   macro avg       0.63      0.63      0.63    220794
weighted avg       0.63      0.63      0.63    220794
```