# DS-GA1007 Final Project User Guide

# Lending Club data Analysis with Visualization

Shangying Jiang (sj2384)

## PROJECT DESCRIPTION

This program provides user the overall view of Lending Club dataset from 2007 to 2011 through Data Visualization and simple machine learning algorithm.

The cleaned dataset has 42445 data instances and 19 features which consist of 10 numerical features and 9 categorical features. This project allow user to study this dataset from several aspects:

- Learn categorical and numerical features separately
- Learn numerical and numerical features separately
- Learn the relationship between numerical and categorical features
- Learn the relationship between one numerical feature and another numerical feature
- Learn the top K most important features that affect the interest rate in a Gradient Boosted Regression Trees model.

## ABOUT DATA

The dataset that used in this project can be found at:
https://www.lendingclub.com/info/download-data.action

The Data Dictionary includes definitions for all the data attributes can also be found at this website.

## BEFORE RUNNING THE PROGRAM

You need

- Python 3.5 or above installed
- Python packages installed:
  - ❖ Required packages: *Pandas, NumPy, matplotlib, seaborn*
  - ❖ All these package can be easily installed with pip in shell/terminal on Windows/Mac:

pip install [The package name]

- Read through this dictionary which includes definitions for all the data features used in this project:

## NUMERICAL FEATURES

| Feature | Definition |
|---|---|
| loan_amnt | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| installment | The monthly payment owed by the borrower if the loan originates. |
| annual_inc | The self-reported annual income provided by the borrower during registration. |
| dti | A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income. |
| delinq_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| inq_last_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| open_acc | The number of open credit lines in the borrower's credit file. |
| pub_rec | Number of derogatory public records |
| revol_bal | Total credit revolving balance |
| total_acc | The total number of credit lines currently in the borrower's credit file |

## CATEGORICAL FEATURES

| Feature | Definition |
|---|---|
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| grade | LC assigned loan grade |
| emp_length | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. |
| home_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| verification_status | Indicates if income was verified by LC, not verified, or if the income source was verified |
| purpose | A category provided by the borrower for the loan request. |
| addr_state | The state provided by the borrower in the loan application |

## HOW TO RUN

Use the terminal/ shell, change the directory where the main function located, and then type:

python main.py

## PROGRAM INPUT

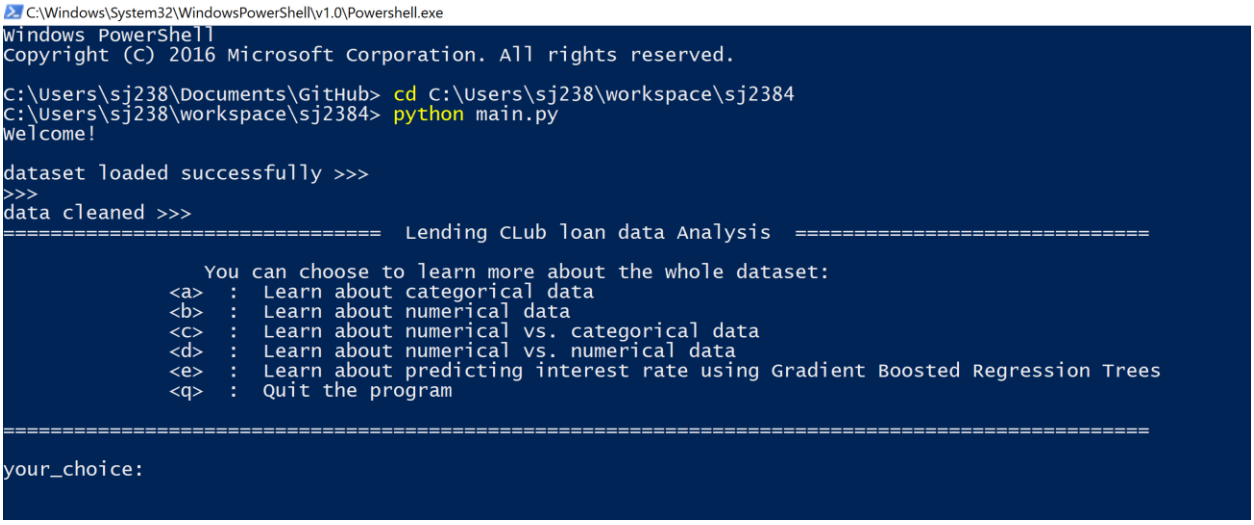
Program takes options inputs like a, b, c etc. from the option list program provided at the beginning of the program. Program also take a keyword input from user when user would like to learn a feature. Program would also ask user how many top important features he/she wants to learn when predicting interest rate.

## PROGRAM OUTPUT

Program could generate different kind of plots of different features upon user's request. In addition, if user want to learn the feature importance in Gradient Boosted Regression Trees model, this program will generate the accuracy of prediction, feature importance of all features and partial dependence of K most important features that specified by user.

## WALK-THROUGH EXAMPLE

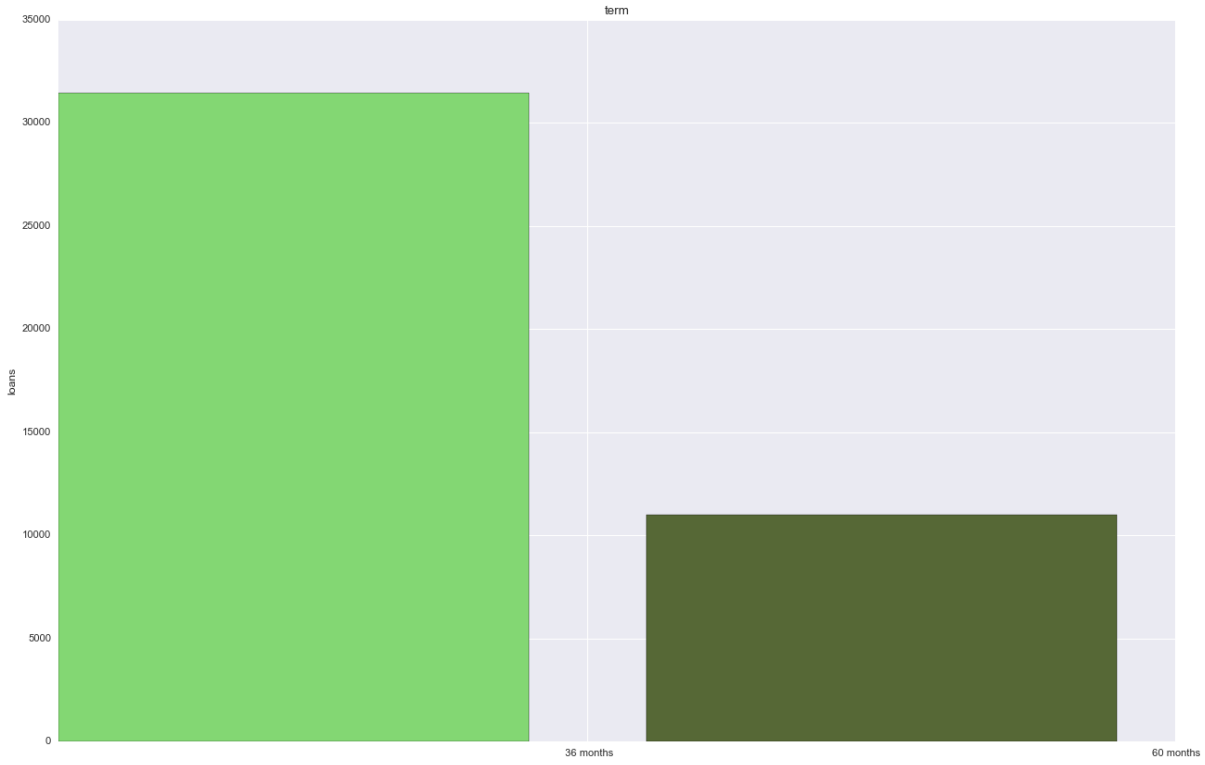
1. At the beginning of the program, you will see an option list:

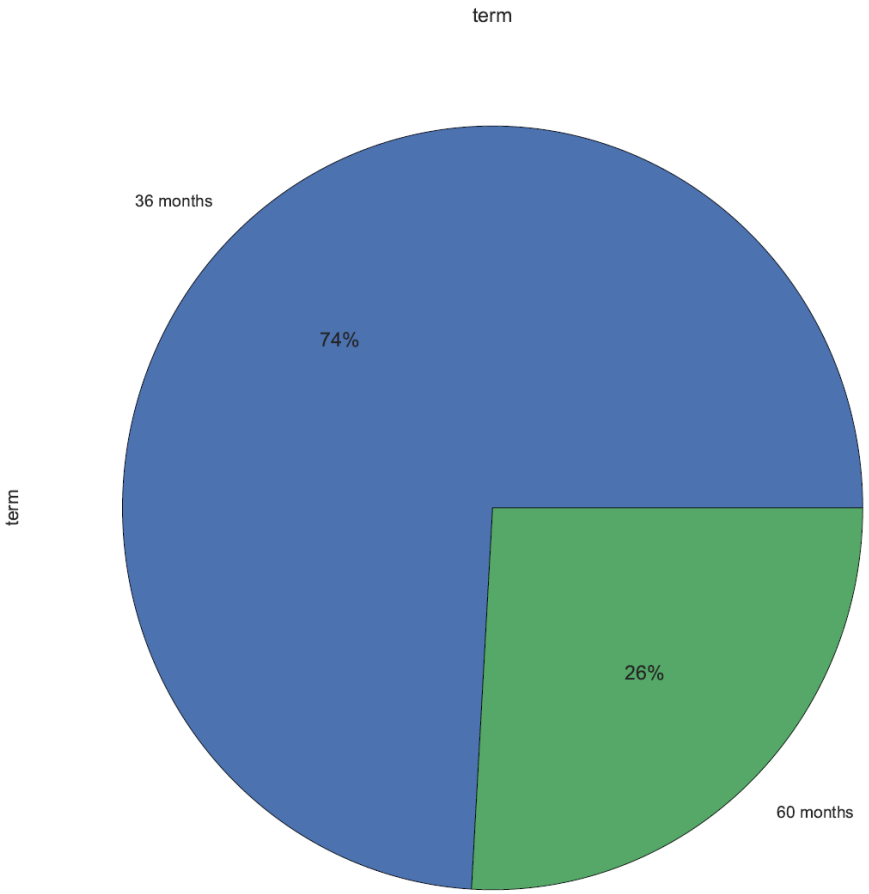



2. Choose one that you are interested in by entering the letter. Suppose you want to learn the categorical data, then you should enter *a*. You will see all the categorical features in the dataset:

```
Windows PowerShell
Copyright (C) 2016 Microsoft Corporation. All rights reserved.

C:\Users\sj238\Documents\GitHub> cd C:\Users\sj238\workspace\sj2384
C:\Users\sj238\workspace\sj2384> python main.py
Welcome!

dataset loaded successfully >>>
>>>
data cleaned >>>
================================  Lending CLub loan data Analysis  ================================

                You can choose to learn more about the whole dataset:
              <a>  :  Learn about categorical data
              <b>  :  Learn about numerical data
              <c>  :  Learn about numerical vs. categorical data
              <d>  :  Learn about numerical vs. numerical data
              <e>  :  Learn about predicting interest rate using Gradient Boosted Regression Trees
              <q>  :  Quit the program

=================================================================================================

your_choice: a
['term' 'grade' 'emp_length' 'home_ownership' 'verification_status'
 'purpose' 'addr_state']
Please select one feature of above you are interested in, or enter q to quit:
```

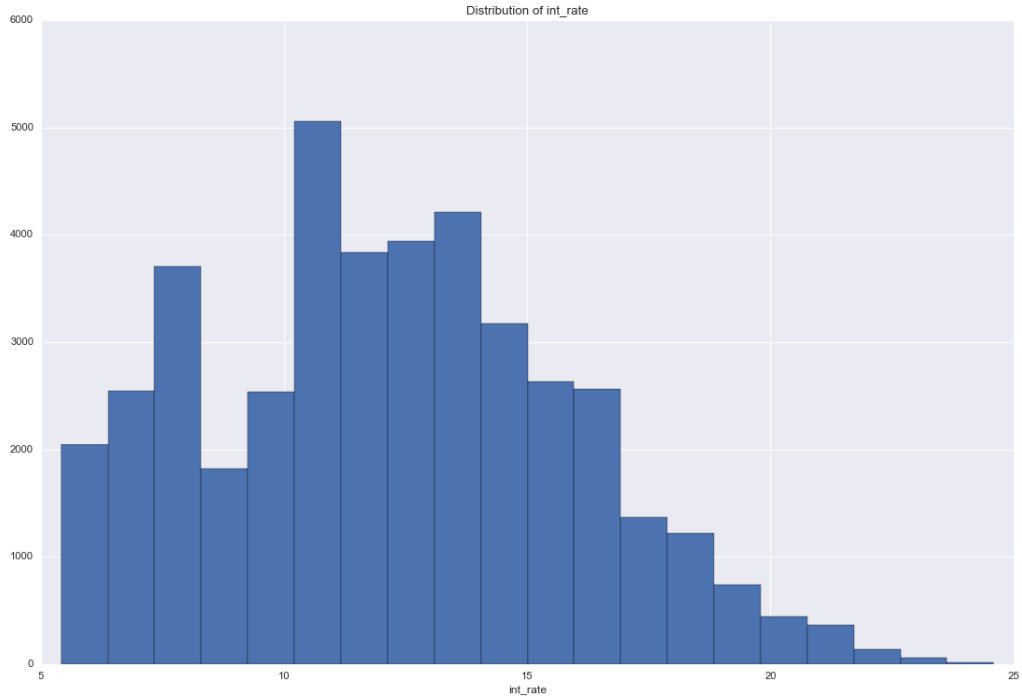
3. Choose one that you are interested in by entering the feature name. Suppose you want to learn the *term*, which is the number of payments on the loan and whose values are in months and can be either 36 or 60. You should enter *term*, then the histogram will pop up, and a pie chart will pop up right after you close the histogram:

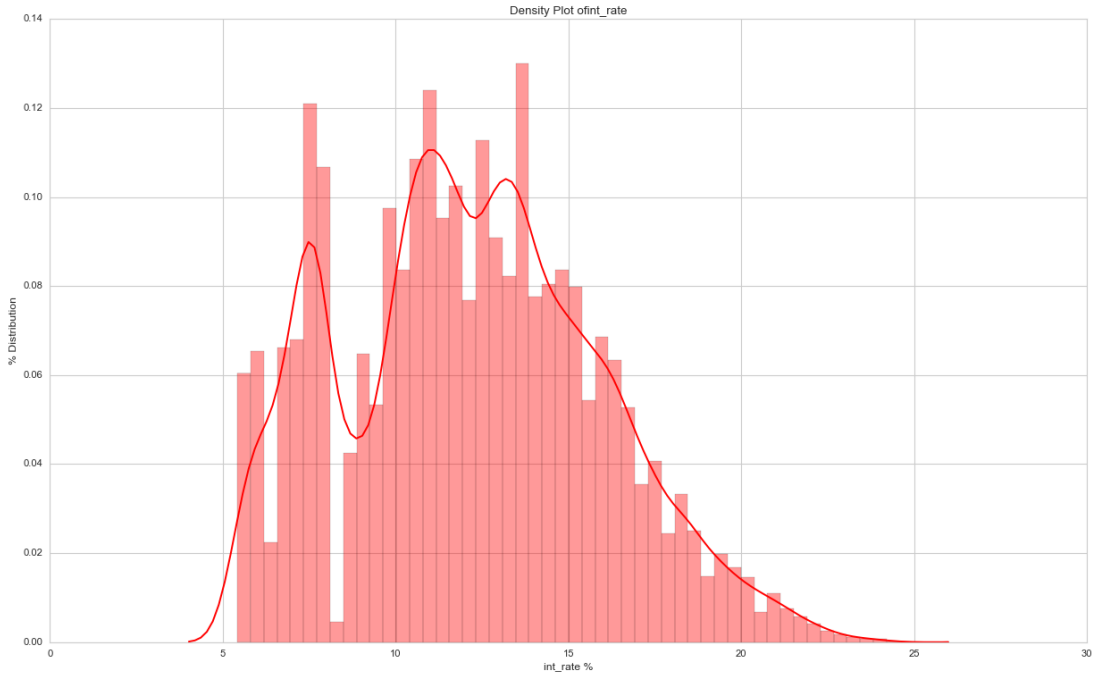




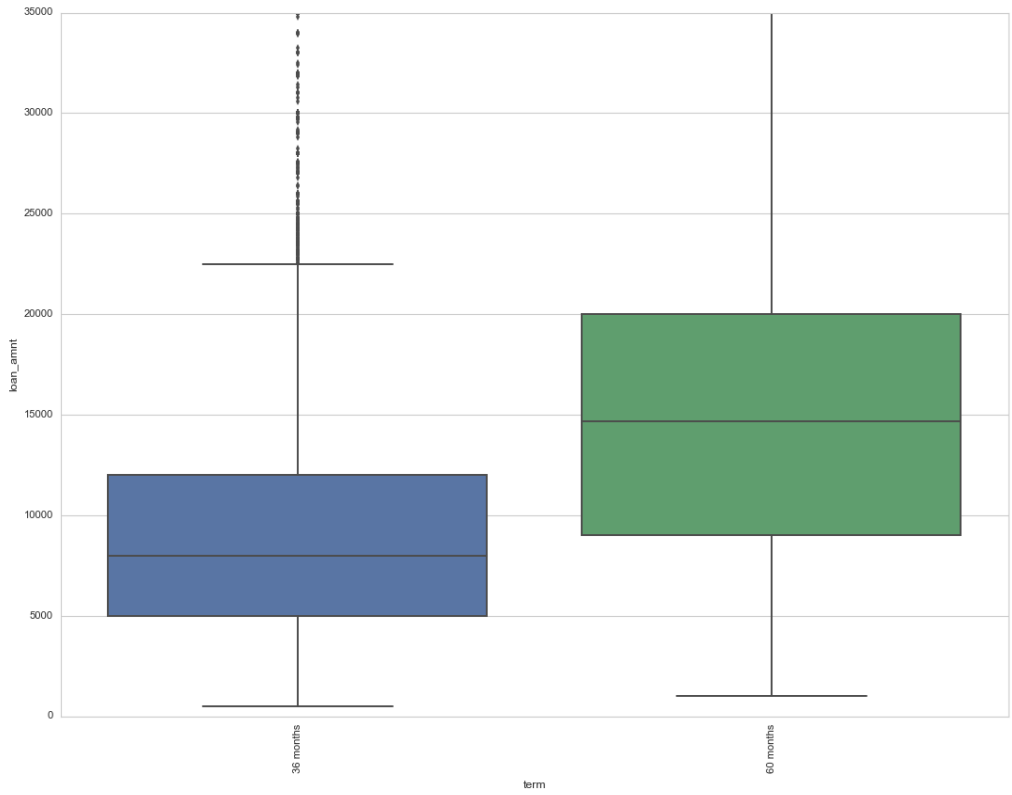

You can close the plot window to go back

4. If you are interested in numerical features, you can enter *b* and you will see all the numerical features in the dataset. Suppose you want to learn about interest rate, you should enter *int_rate*. Then a graph of interest rate's distribution will pop up, followed by the density graph:

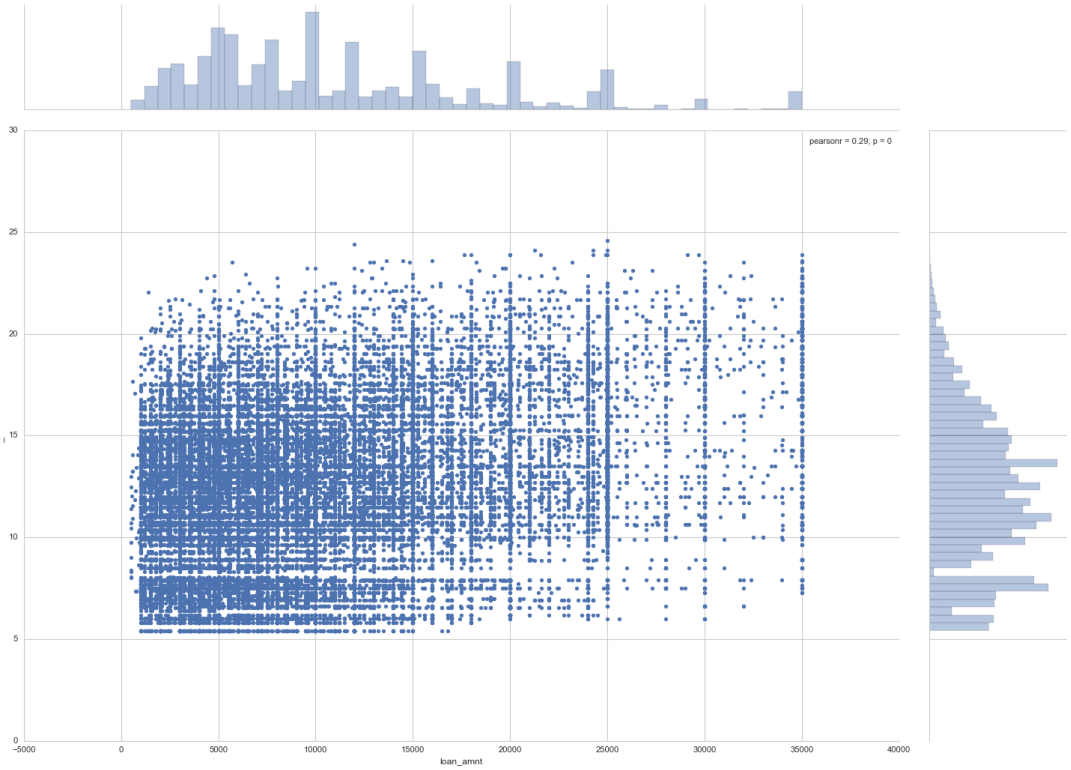

Distribution of int_rate
6000
5000
4000
3000
2000
1000
0
5
10
15
20
25
int_rate


Density Plot ofint_rate
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
% Distribution
0
5
10
15
20
25
30
int_rate %

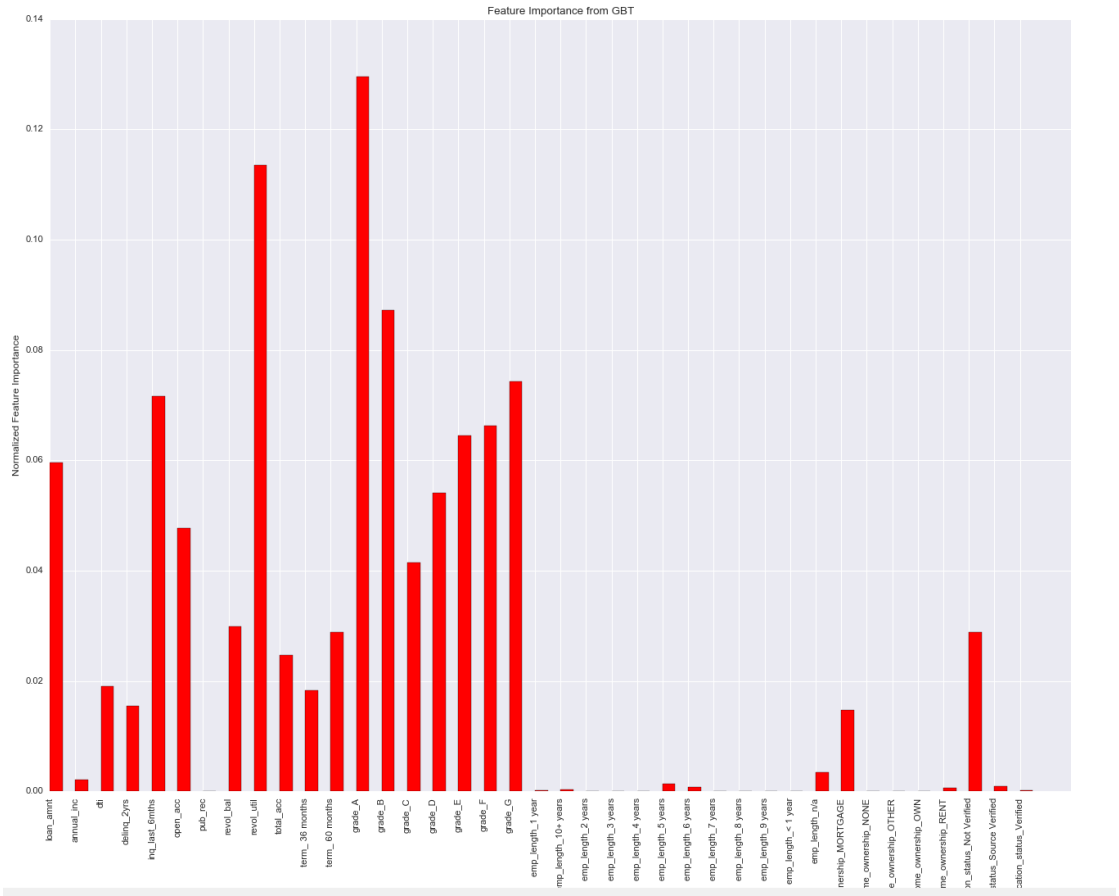
5. If you want to learn numerical feature vs. categorical feature, you can enter *c*, and then enter a categorical feature name followed by a numerical feature name. For example, you want to learn about term vs. loan amount, you should input *term* and then input *loan_amnt*, then you will have a boxplot:

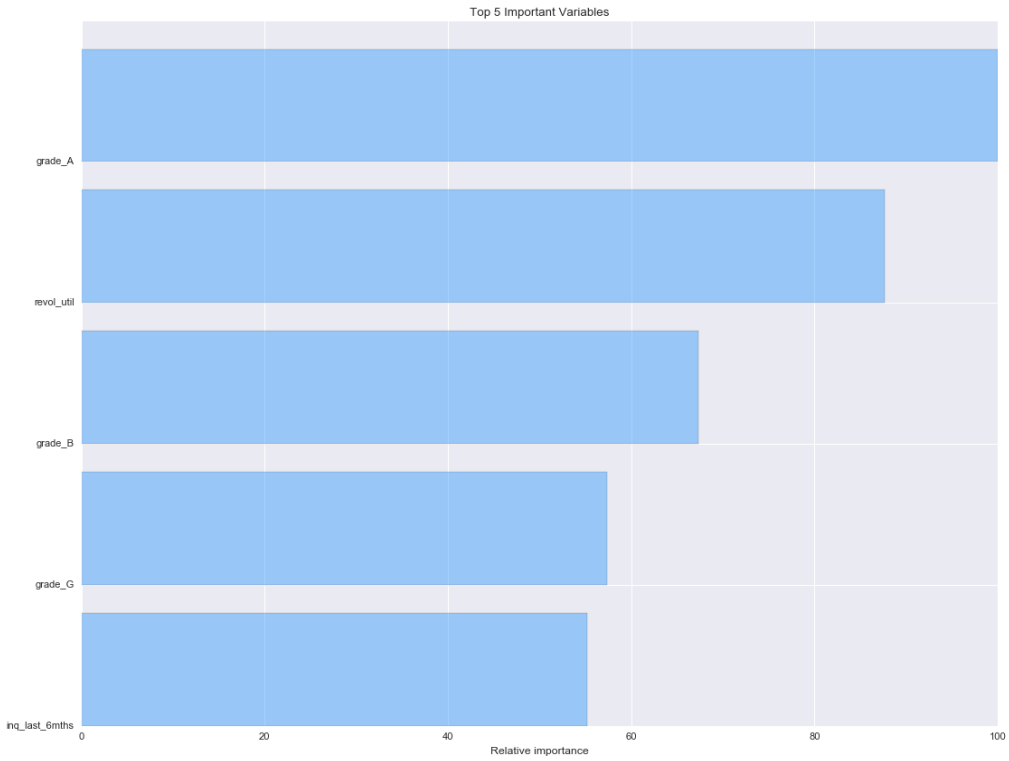


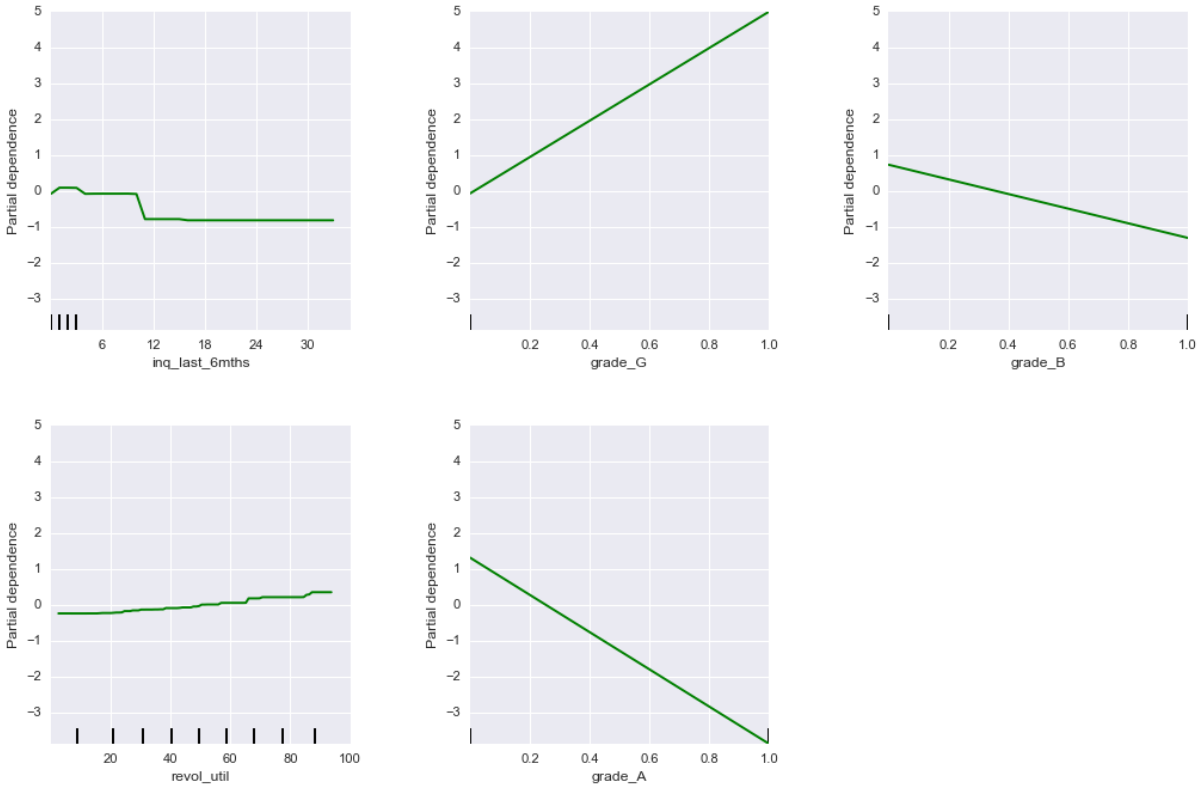

6. If you want to learn numerical feature vs. numerical feature, you can enter *d*, and then enter a numerical feature name followed by another numerical feature name. For example, you want to learn about interest rate vs. loan amount, you should input *iont_rate* and then input *loan_amnt*, then you will have a jointplot:




7. If you enter *e*, you will then be asked to enter a number which is the number of important features you want to learn. For example, if you enter *5*, then you will get a feature importance chart showing importance of all features, a chart showing relative importance of top 5 important features as well as a partial dependence chart of these 5 features:


Feature Importance from GBT
Normalized Feature Importance
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
loan_amnt
annual_inc
dti
delinq_2yrs
inq_last_6mths
open_acc
pub_rec
revol_bal
revol_util
total_acc
term_ 36 months
term_ 60 months
grade_A
grade_B
grade_C
grade_D
grade_E
grade_F
grade_G
emp_length_1 year
mp_length_10+ years
emp_length_2 years
emp_length_3 years
emp_length_4 years
emp_length_5 years
emp_length_6 years
emp_length_7 years
emp_length_8 years
emp_length_9 years
emp_length_< 1 year
emp_length_n/a
nership_MORTGAGE
ne_ownership_NONE
e_ownership_OTHER
me_ownership_OWN
me_ownership_RENT
n_status_Not Verified
tatus_Source Verified
cation_status_Verified


Top 5 Important Variables
grade_A
revol_util
grade_B
grade_G
inq_last_6mths
0
20
40
60
80
100
Relative importance




8. When you decide to end this program, input *q*. Then the whole program will end.