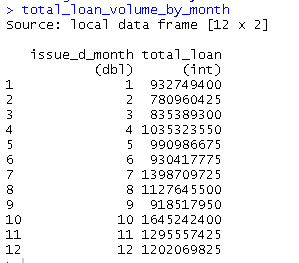
**\*\*All analysis are performed using R programming language (including the query of postgresql AWS). The R code attached along with the submitted documents – inside the zip files\*\***

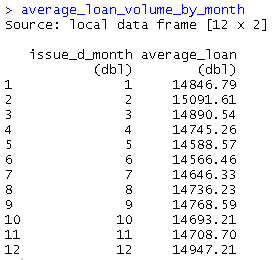
**Set A**

Below are a series of business questions regarding the Lending Club dataset. We need

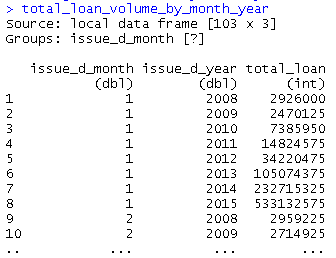
you to determine key performance indicators/metrics (KPIs) that can answer these needs.

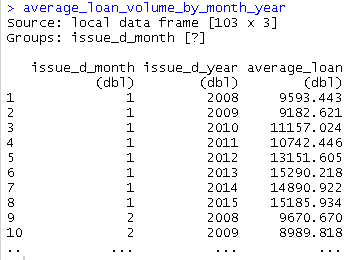
**1. What is the monthly total loan volume by dollars and by average loan size?**





**Comment**: I personally suggest that instead of aggregate everything and group by only month, we should also group by year, thus we can compare how well January 2015 performed against January 2009. Below are some screenshots where I do the same analysis, and group it by month and year:





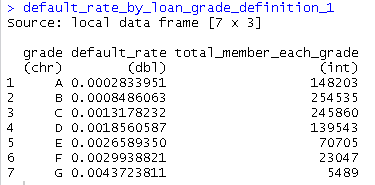
From the table above, we can see how January trend of total and average are compared throughout the years from 2008 to 2015

**2. What are the default rates by Loan Grade?**

According to Investopedia.com(http://www.investopedia.com/terms/d/defaultrate.asp):

There are **two** definitions of default rate:

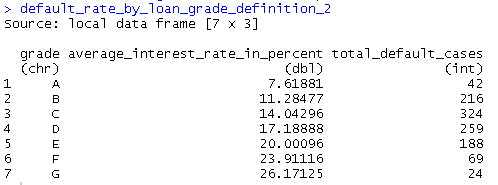
**Definition 1** : The rate of borrowers who fail to remain current on their loans. It is a critical piece of information used by lenders to determine their risk exposure and [economists](http://www.investopedia.com/terms/e/economist.asp) to evaluate the health of the overall economy.



**Comment:** Grade A has the most member, and the total member gradually decreasing from grade A to G. Meanwhile, the default rate is gradually increasing from grade A to G. Statistically speaking, member in group A has probability of .028 % to be default, compared to group G which has probability of .437 % to be default

**Definition 2**: The interest rate charged to a borrower when payments on a revolving [line of credit](http://www.investopedia.com/terms/l/lineofcredit.asp) are overdue. This higher rate is applied to outstanding balances in [arrears](http://www.investopedia.com/terms/a/arrears.asp) in addition to the regular interest charges for the debt.

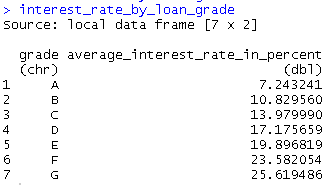
The table below is the average rate of the member who are default, group by each grade



**Comment**:  
This group A has the least average interest percentage, and the interest rate gradually increase from grade A to G. This data slightly misleading since it seems group G only has 24 default cases and seems this group does not deserve to have the highest interest rate. However, keep in mind on the previous discussion we learned that group G has the least members among all group and members of group G are high likely to be default (.437 % chance)

**More insight:**

I tried to calculate the interest rate by the grade (without considering whether the case is default or not), it turned out that group A has the lowest interest, and G has the highest interest. I guess the decision by the company is totally make sense. Members in group G has the highest probability to be default, thus it is necessary for the company to set highest interest rate for this group.



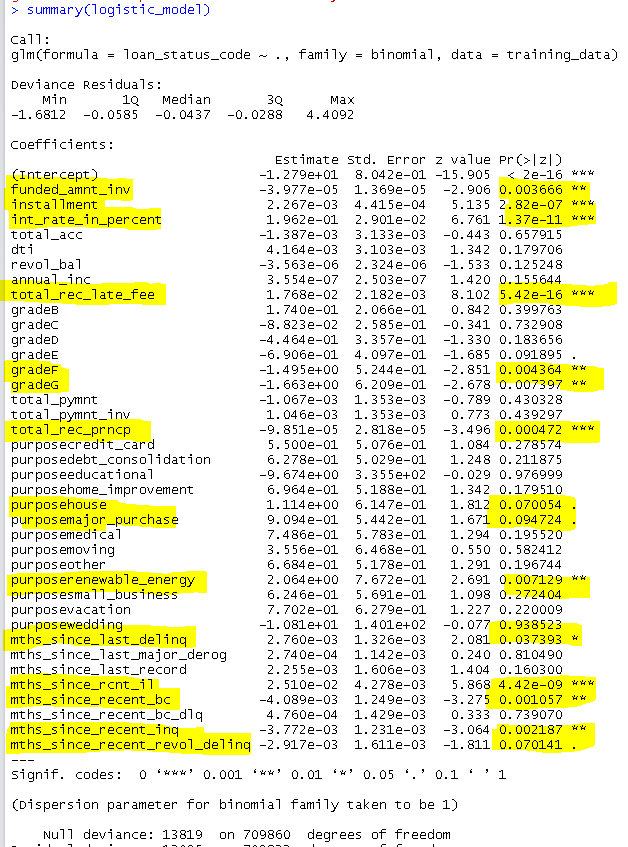
**3. Are we charging an appropriate rate for risk?**

In my opinion: **Yes**, we are charging the appropriate rate for the risk that we calculated. As I previously stated, Members in group G has the highest probability to be default, thus it is necessary for the company to set highest interest rate for this group. Although I agree that group G deserve to have the highest interest rate, I’m not sure if 25.6 percent interest rate is reasonable number for members of group G. I personally think 25.6 % is way too high.

**4. What are the predictors of default rate?**

I run a logistic regression model for this dataset. The output is Default (1) and not Default (0). Below is the summary of logistic model that I performed using stepwise variable selesction. I also test how well this model performed, the proportion training data- testing data is 80% training – 20 % testing

\*\*\*I assume that NAs in month\_since\_(something) variables means that the member has no previous record regarding that particular situation, so I changed those NAs into 120 months (10 years) \*\*\*

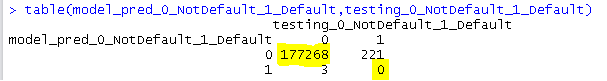


**Comment**:   
According to our logistic regression analysis, the predictor of default case are the variables that has low p-value (highlighted in yellow). They are:

* **Funded\_amnt\_inv** 🡪 The total amount committed by investors for that loan at that point in time.
* **Installment 🡪** The monthly payment owed by the borrower if the loan originates.
* **int\_rate\_in\_percent** 🡪 Interest Rate on the loan
* **total\_rec\_late\_fee 🡪** Late fees received to date
* **Grade 🡪** grade where the member belong
* **total\_rec\_prncp 🡪** Principal received to date
* **Purpose**
* **mths\_since\_last\_delinq 🡪** The number of months since the borrower's last delinquency.
* **mths\_since\_rcnt\_il 🡪** Months since most recent installment accounts opened
* **mths\_since\_recent\_bc 🡪** Months since most recent bankcard account opened.
* **mths\_since\_recent\_inq 🡪** Months since most recent inquiry.
* **mths\_since\_recent\_revol\_delinq 🡪** Months since most recent revolving delinquency.

I also test on how well our model performed. Using 20 percent of testing data, I predict the default and non-default case of testing data using the model that I built, I used the cutoff value of 0.2 of my logistic model to determine the case is Default (labeled as “1”) and non-Default (labelled as” 0”)

below is the confusion matrix:



And also the misclassification error:



It seems that the model performs really well since the misclassification error is very low. However, if we take a look the confusion matrix closer, we can see that our model **predicts really well when the case is not default, but always fail to predict when the case is default**. This happens because out of 887,382 records, only 1122 records are categorized as default (0.126 % of the whole record are default). Since this proportion is relatively tiny compared to the dataset, the model is not well exposed and learn the pattern of the default case. We need more default case in order to build a better model.

**Set B**

**1. Review and QA the dataset and summarize your thoughts on any structural issues:**

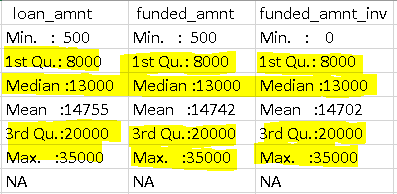
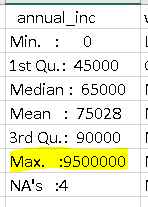
**a. Is there missing data? Is the missing data random or structured: Are some attributes missing more than others?**

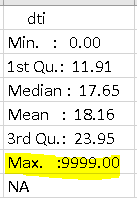
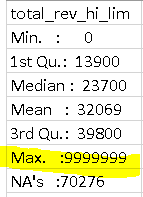
**Yes**, there is a lot of missing data.   
In general, the missing data are structured and have similar pattern (which I guess they all have strong mathematical correlation between each other). More and more missing data on the later columns, compared to earlier columns  
**Yes, some attributes missing more than others,** some of columns with the highest number of missing values are : annual\_inc\_joint, dti\_joint, open\_acc\_6m, open\_il\_6m, etc

\*\*\*I assume that NAs in month\_since\_(something) columns means that the member has no previous record regarding that particular situation, so I changed those NAs into 120 months (10 years) \*\*\*

**b. Are any data values glaringly erroneous?**

**Yes**, some of the values are abnormally high and some of the columns share the same exact maximum values, which is highly unlikely to happen in real world. Some of the examples: (summary statistic in CSV format is attached with the submission):

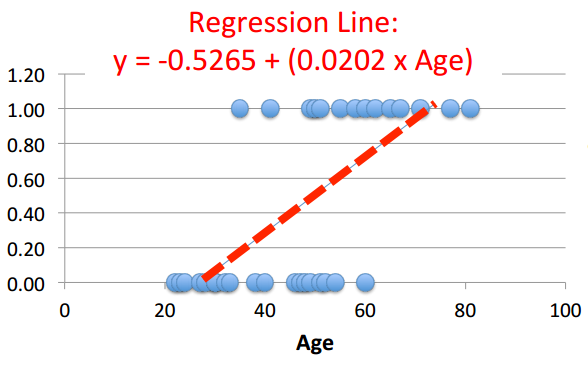
 

**2. Select one of the below topics and CONCISELY explain it to:**

**Topic: LOGISTIC REGRESSION**

1. **Explanation to someone with significant mathematical experience:**

Logistic regression is basically used to handle the limitation of linear regression. One of the main limitation of linear regression is the failure to predict binary format (0 and 1 or Yes and No) of dependent variable. If we plot the linear curve, and input the independent variable(s), the predicted dependent variable will exceed the binary boundary (in this case less than 0 and more than 1) at some values of independent variable(s). One plot that shows this limitation from my previous homework:

  
*\*When we put x=100, the predicted y will exceed 1 (which is not good) \**

We define that the probability of the outcome = 1 given specific value of independent variable(s) as:

Where are the coefficient that we’d like to find. And notice that those probabilities will be bounded between 0 and 1(which is good)

We use that functions and plug in the values from training dataset that we have, and calculate the probability of each record. Then, we take logarithmic function of that probability and calculate the sum of those values. We want to maximize the sum of logarithmic probability, while independently changing the values of . We can use SOLVER feature of Microsoft Excel to find this coefficient. After we come with optimal , we will be able to calculate the probability of situation to be 1 or 0 given specific value of independent variable(s), by inputting that into the “x” of the equation above.

1. **Someone with little mathematical experience:**

Logistic regression is telling you how likely something is going to happen based on different situation.  
Example:  
Let say you want to know if tomorrow is going to be rainy or sunny. You want to predict that based on today’s temperature using logistic regression. When you asked statistician/mathematician to help you decide whether tomorrow you need to bring umbrella, they won’t give you exact answer if tomorrow is going to be rainy or sunny (with 100 percent confident), instead they’re going to tell you probability of the weather being sunny (so you have to bring umbrella just to be safe)

**- Thank you-**