Payoff Intern Assessment

Dataset details: The lending club dataset is a collection of installment loan records, including credit grid data (e.g. FICO, revolving balance, etc.) and loan performance (e.g. loan status).

The data is stored in a postgres database on AWS. Please use the below information to connect to the database with your tool of choice to access the data (R, Python, SQL, etc.)

```
dbname = "intern"
host = "payoff-showtime.ctranyfsb6o1.us-east-1.rds.amazonaws.com"
port = 5432
user = "payoff_intern"
password = 'reallysecure'
```

There are 4 tables for you to use:

- lending_club_2007_2011
- lending_club_2012_2013
- lending_club_2014
- lending_club_2015

Please also use the Lending Club data dictionary as a reference:
https://dl.dropboxusercontent.com/u/1764371/Lending%20Club%20Data%20Dictionary/LCDataDictionary.xlsx

Below are two sections of data questions relating to the Lending Club dataset. Please answer all questions within set A and B.

Quality is MUCH more important than quantity.

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?
2. What are the default rates by Loan Grade?
3. Are we charging an appropriate rate for risk?
4. What are the predictors of default rate?

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?
    b. Are any data values glaringly erroneous?
2. Select one of the below topics and CONCISELY explain it to:
    a. someone with significant mathematical experience
    b. someone with little mathematical experience.
    c. Topics: NoSQL vs SQL, MapReduce, Linear Regression, Logistic Regression, General linear model, Principal Component Analysis, Factor Analysis, K-means Clustering, Support Vector Machines, Markov Process, Hidden Markov Model, Kalman Filter, Decision-tree, Random forest, Kernel density estimation, or the curse of dimensionality.

Important! Please include all code used to query the data and any code you used to generate any analysis or plots in a document and upload it to the link in the included email.