Lending Club Documentation:

Answers:
Part 1:
Categorical Results…
* There are more than 2x as many 36 month loans as 60 month ones.
* The most frequent loan grades in descending order are B > C > A
* The vast majority of loans (in descending order) are Fully Paid > Current > Charged Off
* Loan volume has exponentially increased from 2007 - 2015, and though, still increasing, is somewhat leveling off.
* There is not a significant difference in loan volume by month of origination.

Numerical Results…
* Loan and funded amounts have similar, unimodal, right skewed distributions with long right tails.
(Same Stats Minimum: $500, Maximum: $40,000 — Nearly the same Mean: loan: $14,959.15 / funded: $14,951.84, Std: loan: $9,126.29 / funded: $9,124.39)

* Interest rate has a unimodal, skewed right distribution
(Minimum: 5.31%, Maximum: 30.99%, Mean: 13.07%, Std: 4.81%)

* Annual Income has two very large outliers and otherwise, a unimodal skewed right distribution
(Minimum: $600.00, Maximum: $269,950.00, Mean: $73,940.41, Std: $39,408.32)

* DTI has a somewhat Normal distribution
(Minimum: 0 , Maximum: 0.4 , Mean: 0.18, Std: 0.08)

* Revolving Balance has a unimodal, skewed right distribution with the largest concentration around $5000.
(Minimum: $0.00, Maximum: $88,811.00, Mean: $14,810.26, Std: $13,012.17)

* Total Payment has a unimodal, skewed right distribution with the largest concentration at the lower end of the spectrum.
(Minimum: $0.00, Maximum: $63,296.88, Mean: $11,718.10, Std: $9,765.33)

Part 2:
Question 1: What percentage of loans has been fully paid?

85.9% of 36 month loans with at least 36 months of data were fully paid

Question 2: When bucketed by year of origination and grade, which cohort has the highest rate of defaults? Here you may assume that any loan which was not fully paid had “defaulted”.

Cohort — Year: 2015 / Grade: G — has the highest rate of defaults of 47.0%

Question 3: When bucketed by year of origination and grade, what annualized rate of return have these loans generated on average?

For simplicity, use the following approximation:

Annualized rate of return = (total_pymnt / funded_amnt) ^ (1/3) - 1

Cohort — Year: 2015 / Grade: G — has an annualized rate of return of -1.63%

(vs. Overall ARR without bucketing of 2.66%)

Part 3:

Question 1: Was the model effective?

The model was mostly ineffective at classifying loans. Although, the model was 85.0% accurate in classifying paid vs. unpaid loans, since 85.0% of the loans in the test set were Fully Paid, a similar degree of accuracy could have been obtained by guessing “Fully Paid” across the board.

However, the naive ARR of this blind investment strategy would have attained an ARR of 2.07%. Using the Logistic Regression model with a threshold probability of 0.98, we obtain an ARR of 2.12% — a slight, but noticeable increase over the naive investment strategy.

Question 2: Explain how I validated & describe how I measured the performance of the model?

I segmented the data into 3 buckets — train, cross validation, and test. Because I used month and year of origination as input features, I tried to avoid introducing temporal bias, by using the latest 40% worth of the data as the cross validation and test sets — dividing it randomly and evenly between the two.

Iterating through multiple versions of the model including introducing new features such as total debt (i.e. DTI * annual income) and loan percentage of income (i.e. loan_amount / annual_inc) as well as using one_hot values for the categorical variable “grade”, I identified the best combination of features, using the accuracy values of the cross validation set.

Once, I found the best version of the model, I then calculated the ARR of various portfolios using this model. By using the predicted probably of a loan’s repayment I identified the list of loans in which my model would have been willing to invest, and by finding the version of the model with the highest cross validation ARR, I selected 0.98 as the threshold value for my Logistic Regresion’s investment strategy.

Kaggle Website Data Description: (https://www.kaggle.com/wendykan/lending-club-loan-data)

Update on 2019-03-18 — "Updating files until end of 2018"

Resulting Assumption — This 2007-2018 data represents "present" data as of 11:59 PM on December 31, 2018.
Therefore, any loans originating after Dec. 2015 have less than
36 months of data and should be excluded in Pt. 2 & Pt. 3.

Original Description:

These files contain complete loan data for all loans issued through the 2007-2015,
including the current loan status (Current, Late, Fully Paid, etc.) and latest
payment information. The file containing loan data through the "present" contains
complete loan data for all loans issued through the previous completed calendar quarter.

Assumptions:

##
PART 1: Data Exploration and Evaluation
##
#
Assumptions:
1. Each row is a single unique loan (i.e. there are no duplicate rows for the same loan)
2. Each loan is independent — a default on one does not affect the likelihood of a default on another.
(e.g. multiple loans do not belong to the same person since one of their defaults would increase the
likelihood of a second default.)
3. The loan amount & funded amount represent requested and paid amounts respectively.
4. Column: loan_status — "Does not meet the credit policy. Status:..." can be simplified to just the word(s) after "Status"
5. According to Lending Club on 2007/12/09 — https://blog.lendingclub.com/responsible-lending-better-returns/
"Lending Club maintains very high standards to list a loan, with...a maximum DTI of 30% required."
Thus, assuming that any DTI above 30% is invalid.
6. Lending Club requires income verification — https://help.lendingclub.com/hc/en-us/articles/214502877
Assumption(s): Income is reported prior to origination. Loan recipients must have a non-zero income.

Thus, loans with 0 annual_inc are simply missing data.
7. Column: revol_bal is updated regularly (unlike annual_inc) and therefore cannot be used to verify DTI.

##################################
PART 2: Business Analysis
##################################

Assumptions:
1. (See Above) This 2007-2018 data represents "present" data as of 11:59 PM on December 31, 2018.
Therefore, any loans originating after Dec. 2015 have less than 36 months of data
and should be excluded in Pt. 2 & Pt. 3.
2. Considering only 36 month loans indicates we are not considering a 36 month investment in a 60 month loan. i.e. Holding until maturity.
#
#

Steps Taken:

General Steps:
1. Load Data
2. Create numerical loan grade for logistic regression
3. Separate string date into numerical month / year for pt. 2 & 3
4. Clean up loan_status column by removing “Does not credit policy string”
5. Create combined “lateness” column to simplify creation of “y” column in pt.3 (loan is defaulted simply becomes “lateness” ≥ 1)

Begin Pt. 1:

6. Explore 0s & NULLs in data.
7. Set annual income to NULL if annual_inc ≤ 0. (See Assumption 6.)
8. Since DTI = Monthly Debt / Monthly Income, set DTI to NULL with NULL Income.
9. By definition, DTI is a positive number. Set all negative DTIs to NULL

10. Only 1,725 / 2,260,668 rows have NULL columns. Simply exclude from analysis.

11. Intersperse viewing data plots & cleaning decisions.
- For data with huge outliers (Annual Income / Revolving Balance / DTI), exclude the top 1% of data
- For DTI, many entries seem to be written in percentages as whole numbers, therefore, for everything with a value ≥ 1, divide by 100
- (cont’d) Exclude any DTIs that remain that are above 0.4 (based on Assumption 5).
-

12. Examine data plots. (Following descriptions are post-cleaning.)

Categorical Results…
* There are more than 2x as many 36 month loans as 60 month ones.
* The most frequent loan grades in descending order are B > C > A
* The vast majority of loans (in descending order) are Fully Paid > Current > Charged Off
* Loan volume has exponentially increased from 2007 - 2015, and though, still increasing, is somewhat leveling off.
* There is not a significant difference in loan volume by month of origination.

Numerical Results…
* Loan and funded amounts have similar, unimodal, right skewed distributions with long right tails.
(Same Stats for Minimum: $500, Maximum: $40,000 — Nearly the same for Mean: loan: $14,959.15 / funded: $14,951.84, Std: loan: $9,126.29 / funded: $9,124.39)
* Interest rate has a unimodal, skewed right distribution
(Minimum: 5.31%, Maximum: 30.99%, Mean: 13.07%, Std: 4.81%)
* Annual Income has two very large outliers and otherwise, a unimodal skewed right distribution
(Minimum: $600.00, Maximum: $269,950.00, Mean: $73,940.41, Std: $39,408.32)

* DTI has a somewhat Normal distribution
(Minimum: 0 , Maximum: 0.4 , Mean: 0.18, Std: 0.08)

* Revolving Balance has a unimodal, skewed right distribution with the largest concentration around $5000.
(Minimum: $0.00, Maximum: $88,811.00, Mean: $14,810.26, Std: $13,012.17)

* Total Payment has a unimodal, skewed right distribution with the largest concentration at the lower end of the spectrum.
(Minimum: $0.00, Maximum: $63,296.88, Mean: $11,718.10, Std: $9,765.33)

Begin Pt.2:

13. Load only loans with term contains '36'.
14. Exclude loans with less than 36 months of data (i.e. those that were originated after December, 2015). (See Pt. 2 Assumption 1)

(See code for remaining process and above for answers to questions.)
(Graphs are found within the notebook as well as uploaded separately.)