Please submit your Jupyter notebook with all necessary outputs and comments.

Please download Lending Club's loan origination data from 2007-2015
https://www.kaggle.com/wendykan/lending-club-loan-data
(the dataset (loan.csv) and associated dictionary (LCDataDictionary.xlsx))

For this project, please use the following columns: 'loan_amnt', 'funded_amnt', 'term', 'int_rate', 'grade', 'annual_inc', 'issue_d', 'dti', 'revol_bal', 'total_pymnt', 'loan_status'

Part 1: Data Exploration and Evaluation

Please perform any necessary data engineering (cleaning, aggregation) to understand the dataset. Please point out any data issues you have found (e.g. nulls and outliers) and any assumptions you made to handle it. Describe the data and provide relevant visualization and summary statistics

Part 2: Data Analysis

The goal is to understand if 36 month term loan would be a good investment. Please investigate the following. Assume a 36 month investment period for each loan, and exclude loans with less than 36 months of data available.

1) What percentage of loans has been fully paid?
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".
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

Part 3: Data Modeling

1) The goal is to predict loan default (as defined in question 2 above). What algorithm would you use? Please include your choice and reason in the Jupyter notebook

2) Please build one or more model to predict loan defaults (as defined in question 2 above). Assume that (i) you are given the ability to invest in each loan independently; (ii) you invest immediately following loan origination and hold to maturity (36 months); and (iii) all loan fields that would be known upon origination are made available to you. Was the model effective? Explain how you validated your model and describe how you measure the performance of the model.