Insights into College Debt using Scorecard Data Set

Dr. Amit Nagar

SAP Labs, Palo Alto, CA

3410, Hillview Avenue

Palo Alto, CA 94304

Abstract

The debt burden associated with higher education is a well-known problem. Students incur debt second only to mortgage and continue to carry it long after graduating. Their ability to repay is dependent upon completion, their future earning power and other financial commitments. The students who drop out tend to default. In either case, debt acquired in process of getting higher education stands to impact future financial wellness of the students.

Now document the key observations…\*\*\*

Keywords: College Debt, Data Science, R

Insights into College Debt using Scorecard Data Set

This analysis uses the data sets published by the US Department of Education (DoE). As per the published documentation, it consists of the data collected from 1996 to 2013 and sourced from

* federal reporting by institutions,
* federal financial aid, and
* tax information.

In this paper, we focus on 2013 data which is stored in file *merged2013\_pp.csv*. The data set has a dimension of 7804 (rows) by 1729 (columns). It is divided into sub-sections as shown in Table 1. Each sub section has a number of attributes. We will use the sub section based demarcation in subsequent analysis in this paper.

**Table 1**: Scorecard Categories

| Data | Description |
| --- | --- |
| Root | technical information |
| About the school | information about the school |
| Academics | academic offerings |
| Admission | Scores |
| Completion Statistics | completion rates |
| Cost | cost to students |
| Earnings | earnings & employment prospects |
| Financial Aid | federal financial aid, debt ceiling and loan performance |
| Outcomes for Title IV students | completion rates that track institutional outcomes for students |
| Earnings | earnings of former students |
| Repayments | debt repayment |

Our focus in this analysis is on the **repayment rates**. As defined in the data set documentation, repayment rates “depict the fraction of borrowers at an institution who have not defaulted on their federal loans and who are making progress in paying them down (i.e. have paid down at least $1 in the principal balance on their loans) after leaving school (RPY\_\*YR\_RT). The rates are available for 1 (\_1YR\_RT), 3 (\_3YR\_RT), 5 (\_5YR\_RT), and 7 (\_7YR\_RT) years after leaving school."

To further narrow down we focus on the 7-year repayment rate, namely, **RPY\_7YR\_RT** (repayment rate). We will like to understand how repayment rate is affected by the other columns in this data set. In other words, we formulate this as an inference problem with repayment rate as the dependent or response variable and the remainder of the variables as independent or explanatory variables. We feel that repayment rate is a key metric that will be of interest to decision makers. Of course, 3-year and 5-year statistics will likely provide an early indication on the health of the debt repayment. We will make relevant observations regarding the temporal performance of the repayment statistic where relevant

# Understanding the Data Set

Examining the data in the csv file, we notice a number of data points marked as NULL or Privacy Suppressed as shown in figure 1.

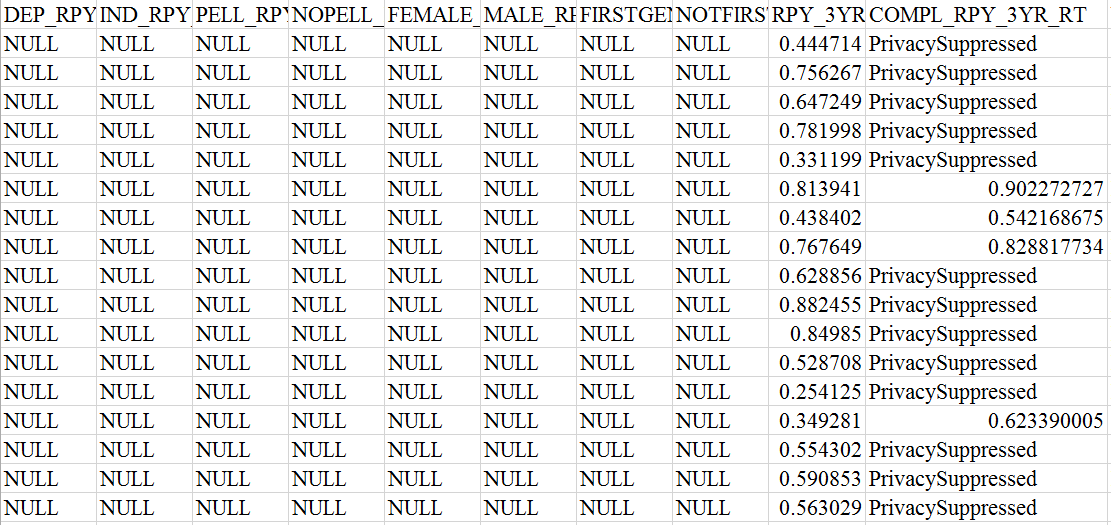


Figure 1. CSV file snapshot showing NULL and Privacy Suppressed data points

These constitute the missing data in the data set. We will like to understand the extent of this missing data. In R, this can be accomplished if we convert the data marked NULL or Privacy Suppressed to NA. NA is a (vector) data type in R supported by language features that allow for special processing. We achieve this when reading the data from the csv file using the following command:

**if** ( !( "scorecard" %in% ls() )){

scorecard = (read.csv(datafiles,header=TRUE, sep=",", stringsAsFactors = FALSE,strip.white = TRUE, na.strings = c("NULL","PrivacySuppressed

")))

}

We first examine the prevalence of the NAs in the rows of the data set. R provides a facility to count the number of NAs in rows of a data set using the *complete.cases* command. As used below, *complete.cases* returns a logical vector indicating which vectors are not complete or have missing values.

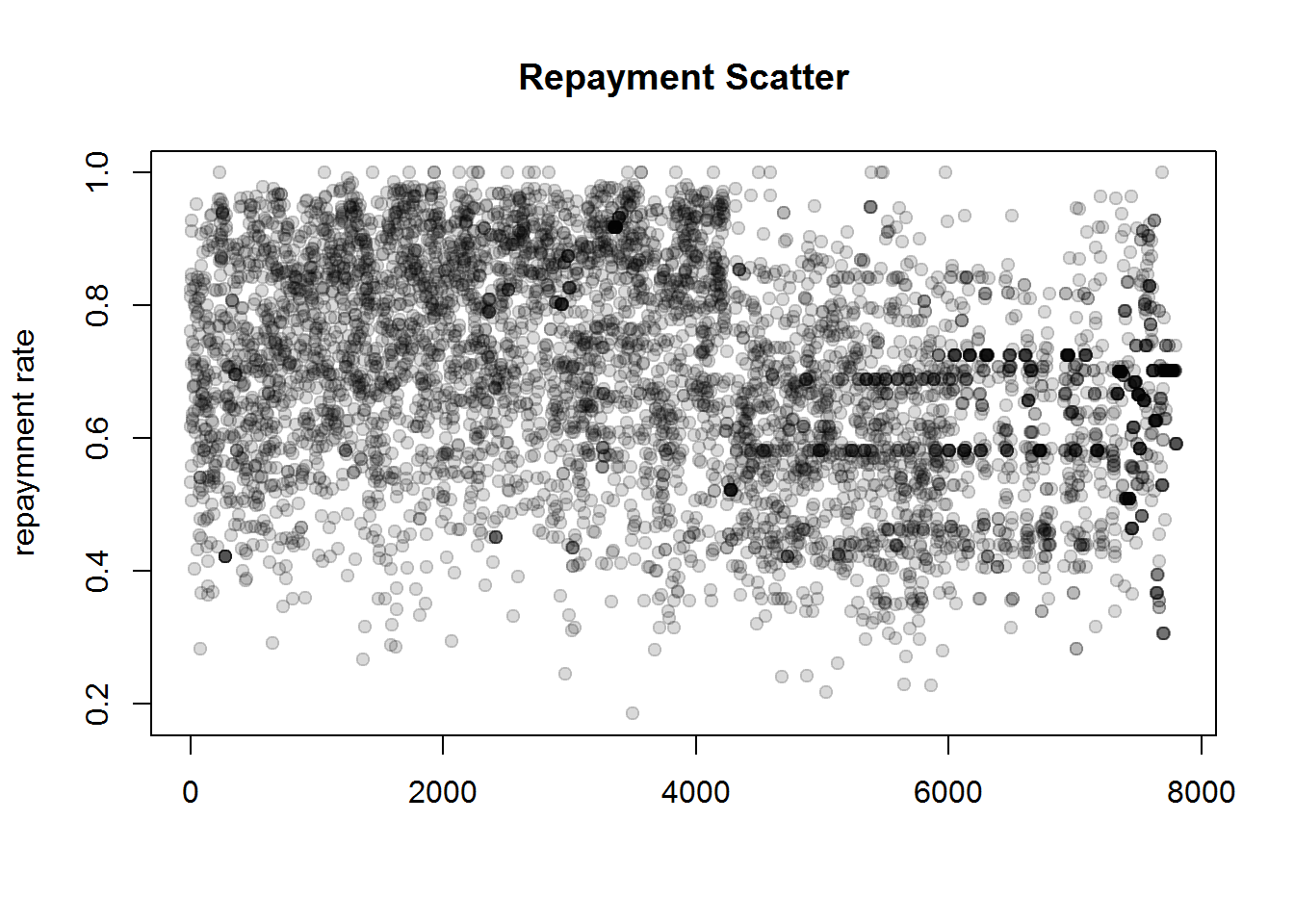
nrow(scorecard[!complete.cases(scorecard),])

The number of rows returned by above is 7804. This means that each row of the scorecard data set has at least one missing value. In fact, minimum number of NAs in the rows of the data set is 70.10% and maximum is 98.96%.

Examining, columns for missing values is even more revealing. In this data set, there are 1174 (67.4%) columns in which all values are missing (or NA). We believe in such cases, imputation is neither feasible nor a useful strategy. In other words, we cannot substitute NAs for any meaningful value. Therefore, we remove all the 1174 NA columns from further consideration from the data set. The reduced data set has a dimension of 7804 (rows) by 555 (columns). The dimension from a column perspective is reduced considerably without loss of any information.

## Exploring Repayment Patterns

Next let us do a visual exploration of our independent variable, repayment rate. Figure 2 depicts scatter plots of entire scorecard data set. Our aim is to detect any patterns that might be readily visible in the scatter plot.



**Figure** 2. Scatter plot of Repayment Rates

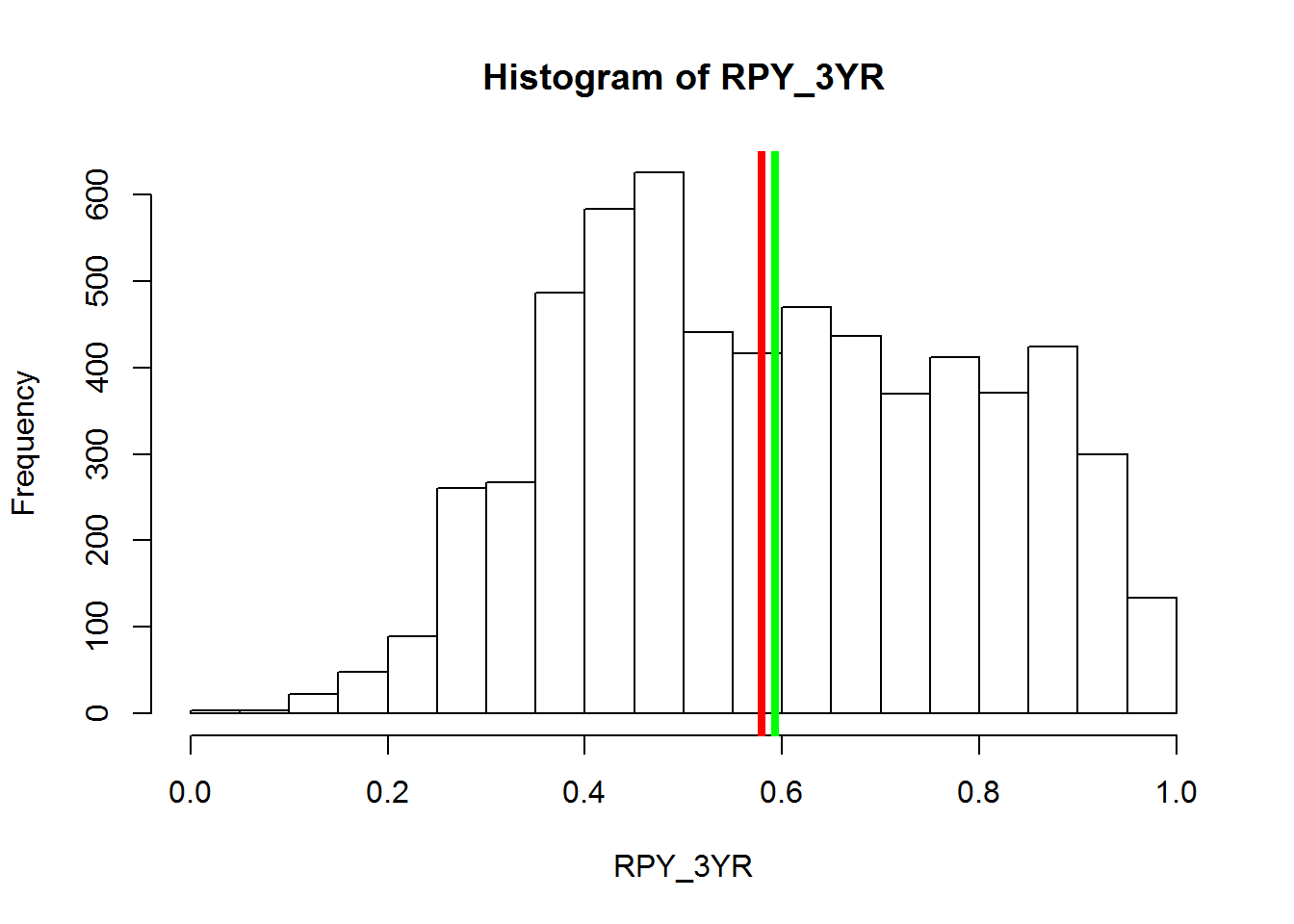
There are no discernible patterns in the plots which provide any indication of the shape of underlying distribution. It does provides some interesting insights:

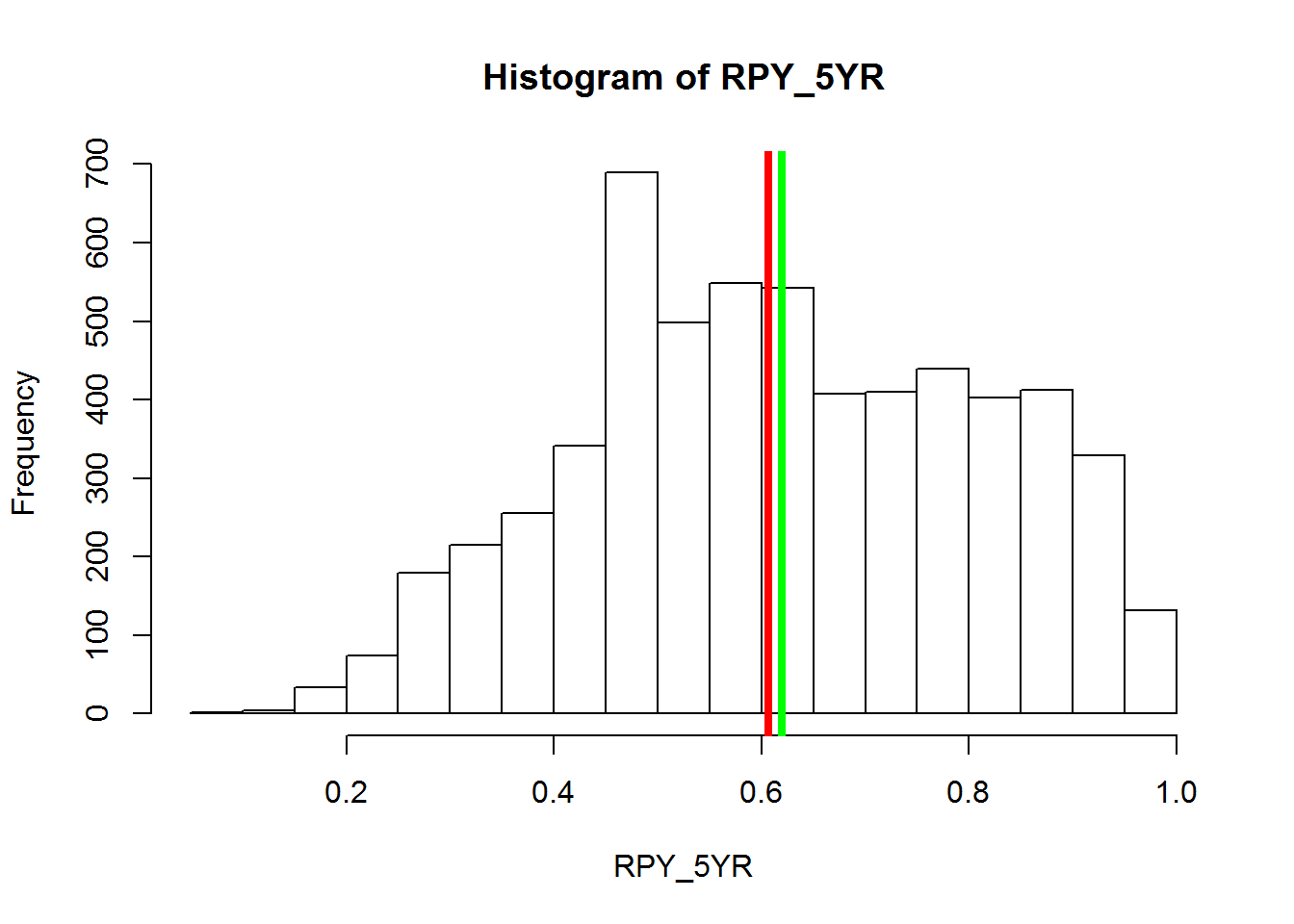
* 20-30 institution have a perfect repayment rate of 1.0. Are there any learnings in this data to understand what these institutions might be doing different?
* an interesting pattern that seems to emerge is around clustering of the repayment rates
  + [0.7, 1.0] for first 4000 data points , and
  + Around 0.6 for the data points from data point 4000 onwards. This is quite an interesting pattern as we have not sorted the data set on repayment rates
* relatively fewer institutions have repayment rates under 0.4. Is there an underlying reason for this pattern?

Across the Years.

We notice a couple of interesting trends if we compare the repayment rate across 3, 5, and 7 years.

* Increasingly prominent negative skew indicating a clustering of data towards larger repayment rates as time progresses,
* In addition, we also notice that the mean (green line) and median (red line) cross over with median becoming larger than the mean value at 7th year mark. This is in agreement with the increasing negative skew that we alluded to in the first point above.





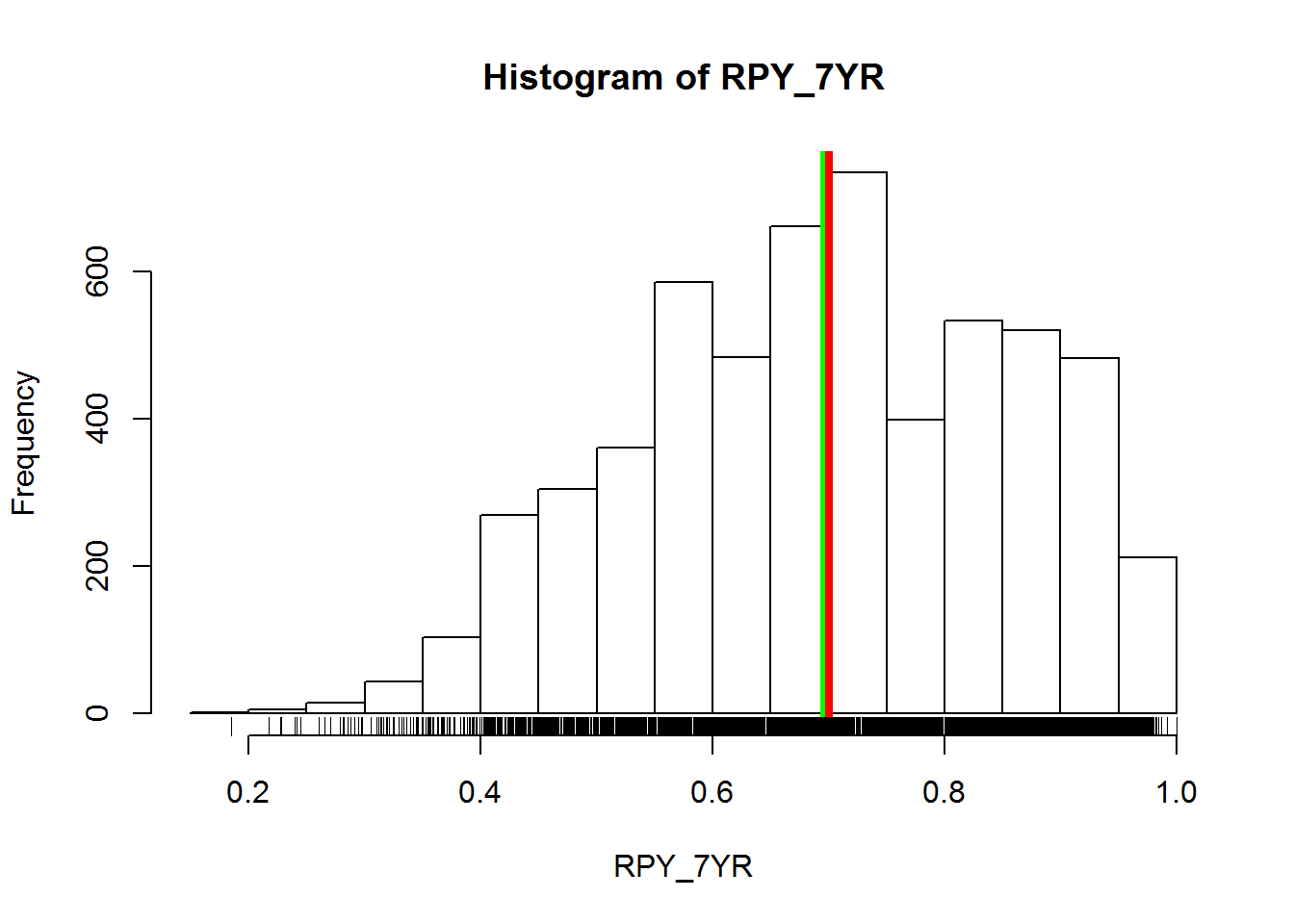


Figure 3 Repayment Rates across years

References

[1] <https://www.kaggle.com/c/us-dept-of-education-college-scorecard>

[2] [http://stats.stackexchange.com/questions/28576/filling-nas-in-a-dataset-with-column-medians-in-r](http://stats.stackexchange.com/questions/28576/filling-nas-in-a-dataset-with-column)

[3] <http://www.stat.berkeley.edu/~s133/Cluster2a.html>

[4] Peng, Roger (2015). Exploratory Data Analysis with R. LeanPub.

Footnotes

1

Tables

Figures