**Handling Missing Data**

1. **Columns Filled with 0 are below:**
   1. **delinq\_2yrs** - Blank values over here would mean that the borrower was not delinquent in past 2 yrs. Hence replaced with value 0 for no delinquency.
   2. **inq\_last\_6mths** - Missing values would mean that there won’t have been any inquiries in past 6 months. Hence filled with 0.
   3. **mths\_since\_last\_record** - This column is filled with value 0 because no value over here would mean that the the dataset has no information about last record. Hence filled with 0.
   4. **open\_acc** - Filled with 0 because empty cell would mean that borrower has closed all of its credit lines.
   5. **pub\_rec** - Blank cell over here would mean that no public record found on paper regarding the borrower. Thus filled with 0.
   6. **total\_acc** - blank value over here would mean that borrower has closed all accounts. Hence replaced with 0.
   7. **collections\_12\_mths\_ex\_med** - Significantly less number of blank values. Hence replace with 0.
   8. **mths\_since\_last\_major\_derog** - Empty value here would infer that the borrower won’t have any worse rating on his account. Hence for consistency replaced with 0.
   9. **annual\_inc\_joint** - Replaced with 0.Since significantly less number of blank values.
   10. **dti\_joint** - Blank values would mean that the account is not a joint account. Hence for consistency replaced with 0.
   11. **acc\_now\_delinq** - Blank values would mean that account has never been delinquent. Thus replaced with 0.
   12. open\_il\_6m - Blank value here would mean that currently no account is being opened by the borrower. Hence replaced with 0.
   13. **acc\_open\_pas t\_24mths** - Blank value here would mean no account has been opened past 24 months. Thus replaced with 0.
   14. **chargeoff\_within\_12\_mths** - Blank value here would mean no chargeoff has been posted on account for the past 24 months. Thus replaced with 0.
   15. **delinq\_amnt** - blank would mean no delinquency hence replace with 0 i.e. Delinquent amount.
       1. All the months since account below are replaced 0 because empty value here would have meant the borrower has no record for the specified variable.
          1. mo\_sin\_old\_il\_acct
          2. mo\_sin\_old\_rev\_tl\_op
          3. mo\_sin\_rcnt\_rev\_tl\_op
          4. mo\_sin\_rcnt\_tl
          5. mths\_since\_recent\_inq
          6. mths\_since\_recent\_bc
          7. mths\_since\_recent\_revol\_delinq
   16. **mort\_acc** - Blank value infers that no mortgage account has been opened. Hence replaced with 0.
   17. **num\_accts\_ever\_120\_pd** - This column describes the number of accounts ever past 120 days due. Thus blank here should be replaced by 0 since there has been no account for the borrower that is being due.
   18. All the below variables are **replaced with 0** since the borrower won’t have that particular account.
       1. num\_actv\_bc\_tl
       2. num\_actv\_rev\_tl
       3. num\_bc\_sats
       4. num\_bc\_tl
       5. num\_il\_tl
       6. num\_op\_rev\_tl
       7. num\_rev\_accts
       8. num\_op\_rev\_tl
       9. num\_rev\_tl\_bal\_gt\_0
   19. **num\_tl\_120dpd\_2m** - Number of accounts currently past 120 days due would be blank if none of his accounts have been past 120 days ever.
   20. **num\_tl\_30dpd** - Same as num\_tl\_120dpd\_2m. Number of accounts currently past 30 days is also replaced with 0.
   21. **num\_tl\_90g\_dpd\_24m** - Number of accounts 90 or more days past due in last 24 months also replaced with 0 since for the same reason being as above that account has never been past due.
   22. **num\_tl\_op\_past\_12m** - Blank value here would have meant that no accounts has been opened by the borrower in past 12 months. Hence replaced with 0 to indicate 0 accounts.
   23. **pub\_rec\_bankruptcies** - Blank value here would be an obvioues guess to replace by 0. Since any other value would have a major impact on the borrower’s account.
   24. **tax\_liens** - For the same reason as public recorded bankruptcies. This is also filled with 0.
2. **Columns Filled with Median are below:**
   1. **annual\_inc** - Replacing blank with median would be the obvious choice here since the data includes income of the borrower and that could be a deciding factor to predict interest rate. Thus replacing with median would keep the data consistent.
   2. **revol\_util** - Large number of empty fields hence replaced with median.
   3. **tot\_coll\_amt** - Large proportion of empty data thus replaced with median.
   4. Below are remaining columns that have been replaced with median for the same reason of large amount of empty cells:
      1. tot\_cur\_bal -
      2. open\_acc\_6m
      3. avg\_cur\_bal
      4. bc\_open\_to\_buy
      5. pct\_tl\_nvr\_dlq
      6. percent\_bc\_gt\_75
      7. tot\_hi\_cred\_lim
      8. total\_bal\_ex\_mort
      9. total\_bc\_limit
      10. total\_il\_high\_credit\_limit
3. **Columns filled with constant values:**

Below are some of the columns whose blank values are replaced with constant values in consistency with the other values.

* + 1. **emp\_title**
    2. **Desc**
    3. **Title**

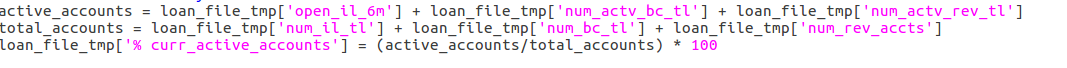
1. **Columns dropped while handling missing data:**

The following columns have been dropped in the Handling Missing Data stage itself due to large proportion of empty values and also they won’t play further part in predicting interest rates.

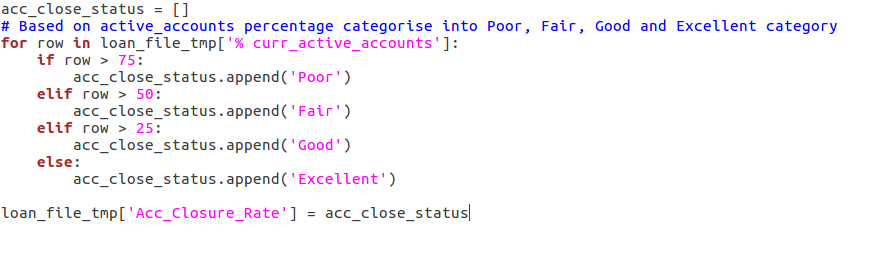
* 1. **mths\_since\_rcnt\_il** - Months since most recent installment accounts opened
  2. **total\_bal\_il** - Total current balance of all installment accounts
  3. **il\_util** - Ratio of total current balance to high credit/credit limit on all install account.
  4. **open\_rv\_12m** - Number of revolving trades opened in past 12 months
  5. **open\_rv\_24m** - Number of revolving trades opened in past 24 months
  6. **max\_bal\_bc** - Maximum current balance owed on all revolving accounts
  7. **all\_util** - Balance to credit limit on all trades
  8. **inq\_fi** - Number of personal finance inquiries
  9. **total\_cu\_tl** - Number of finance trades
  10. **inq\_last\_12m** - Number of credit inquiries in past 12 months
  11. **bc\_util** - Ratio of total current balance to high credit/credit limit for all bankcard accounts.
  12. **mths\_since\_recent\_bc\_dlq** - Months since most recent bankcard delinquency
  13. **last\_pymnt\_d** - Last month payment was received
  14. **next\_pymnt\_d** - Next scheduled payment date
  15. **Last\_credit\_pull\_d** - The most recent month LC pulled credit for this loan.

**Feature Engineering :**

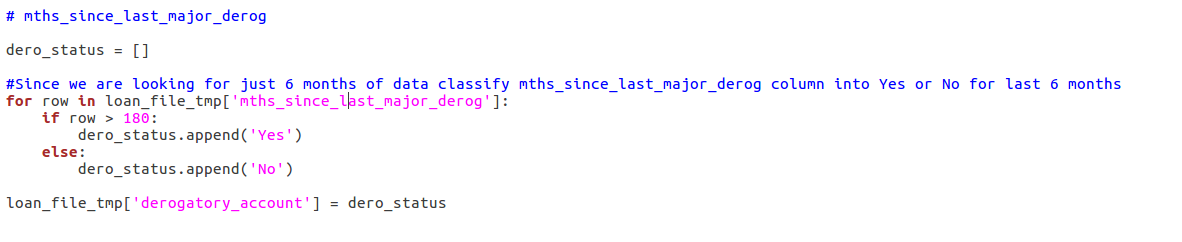
1. We have calculated loan performance for the each and every individual based on some paramteres below:
   1. Calculated the total number of accounts(Installment accounts, Bankcard Accounts, Revolving Accounts) a borrower has ever opened.(say T)
   2. Calculated Total number of currently active accounts.(say A)
   3. % of currently active accounts = T/A
   4. This percentage value is then compared with some threshold values and categorised into a new column called Account Closure Rate(Acc\_Closure\_Rate) how the Loan Closing performance is for a particular member(Poor, Fair, Good, Excellent).
   5. This coulmn would play a major part in preicting interest rates for a person based on his Loan Performance.
   6. Code to calculate percentage.



* 1. Categorise:

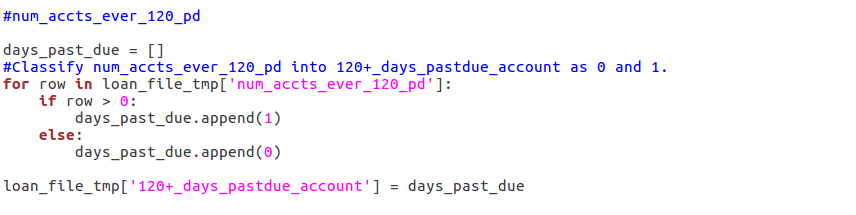


1. Since we are giving importance on the last 6 months of record to predict interest rates. The Column, Months since most recent 90-day or worse rating(mths\_since\_last\_major\_derog) can be derived into a new column called Derogatory Account(derogatory\_account). This new column is derived as below:



For the new column we are just considering months greater than 180 and then categorising it into Yes if the account is derogatory or No if it is not.

1. Another derived column would be 120+ days past due account. This column would derive whether the account has ever been 120+ past days due or not. This is derived from the column named number of accounts ever 120 days past due(num\_accts\_ever\_120\_pd)



1. Dropped Columns:
   1. Some of the variables such as month, date, desc, title, zipcode etc. That are not significant in calculation interest rates are dropped in feature engineering. Detailed list of such columns are:
   2. id,member\_id,installment,grade,sub\_grade,issue\_d,loan\_status,pymnt\_plan,url,desc,title,zip\_code,addr\_state,earliest\_cr\_line,inq\_last\_6mths,mths\_since\_last\_record,initial\_list\_status,out\_prncp,out\_prncp\_inv,total\_pymnt,total\_pymnt\_inv,total\_rec\_prncp,total\_rec\_int,total\_rec\_late\_fee,recoveries,collection\_recovery\_fee,collections\_12\_mths\_ex\_med,policy\_code,tot\_coll\_amt,tot\_cur\_bal,open\_acc\_6m,delinq\_amnt,mo\_sin\_old\_il\_acct,mo\_sin\_old\_rev\_tl\_op,mo\_sin\_rcnt\_rev\_tl\_op,mo\_sin\_rcnt\_tl,mths\_since\_recent\_bc,mths\_since\_recent\_inq,mths\_since\_recent\_revol\_delinq,num\_accts\_ever\_120\_pd,num\_actv\_bc\_tl,num\_actv\_rev\_tl,num\_bc\_sats,num\_bc\_tl,num\_il\_tl,num\_op\_rev\_tl,num\_rev\_accts,num\_rev\_tl\_bal\_gt\_0,num\_sats,num\_tl\_120dpd\_2m,num\_tl\_30dpd,num\_tl\_90g\_dpd\_24m,num\_tl\_op\_past\_12m,pub\_rec\_bankruptcies,tax\_liens,% curr\_active\_accounts,mths\_since\_last\_major\_derog

**LUIGI PIPELINE:**

**Task 1: DownloadData()**

* + - 1. **This task downloads data from the given url and merges them into a single file in the output.**

**Task 2: CleanData()**

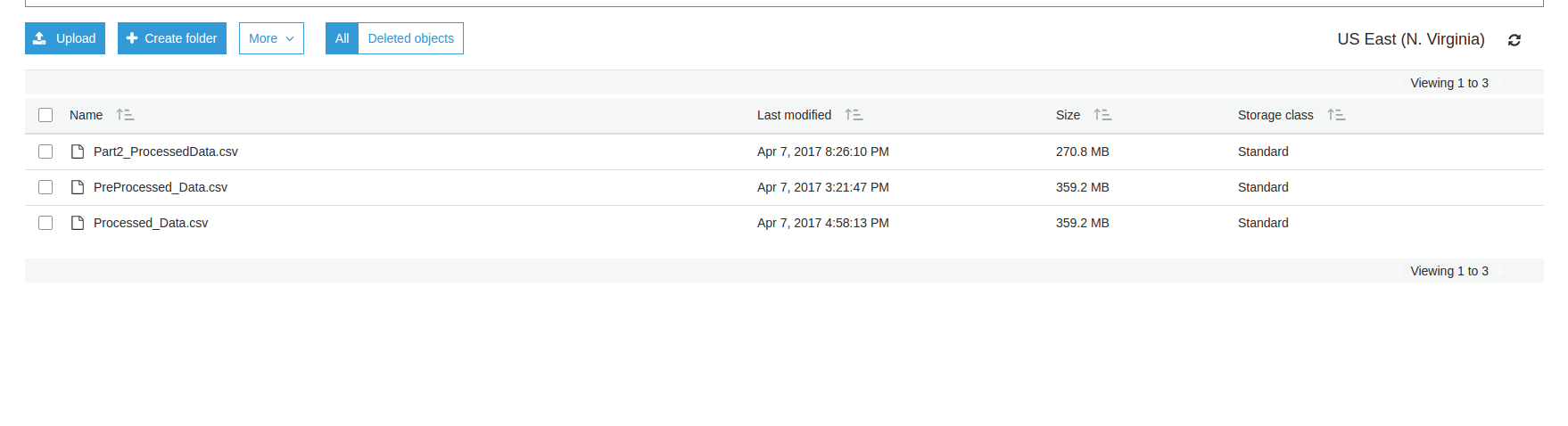
* + - 1. **This task handles missing data as described above and cleans the merged file received as an input from the output of the Task1.**

**Task 3: FeatureSelection()**

* + - 1. **This task takes the cleaned file as an input from the dependent task and does variable selection on some parameters described above and derives some new columns while deleting old one’s.**

**Task 4: UploadToS3()**

* + - 1. **This task uploads the output file received after pre processing in above task on the Amazon S3 bucket.(Processed\_Data.csv)**



**REJECTED LOAN DATA**

**HANDLING MISSING DATA & FEATURE ENGINEERING**

1. **Columns Filled with Constant are below:** 
   1. **Loan Title** - Blank Values are filled with Loan Title Not Specified.
   2. **State** - Very less number of empty fields in this column compared to the total number of records. Thus replaced with XX that would not affect the entire data.
2. **Application Date** - This field is derived into Year, Month and Date columns.
3. **Risk\_Score -** This column is very usefull in analysis. Hence we have taken only what is available for this column while discarding all the emtpy rows.
4. **Employment Length -** this column is pre-processed to have values 0,1,....,10 for years instead of >1 year and 10+years.

**LUIGI PIPELINE:**

**TASK 1: DownloadData()**

* + - 1. **This task downloads data from the given url and merges them into a single file in the output.**

**TASK 2: CleanData()**

* + - 1. **This task handles missing data and variable selection as shown above.**

**TASK 3: UploadToS3()**

* + - 1. **This task uploads the output file received after pre processing in above task on the Amazon S3 bucket.(Pre\_Processed\_Data.csv)**

