Data Extraction:

Downloaded the dataset in CSV format (loan.csv) and associated dictionary (LCDataDictionary.xlsx) from <https://www.kaggle.com/puneeshk/lending-loan-club-dataset>.

*Data Exploration and Evaluation:*

Cleaning:

* Cleaning Na columns by dropping them from original data set created as Loan\_Df
* Dropping columns That has same values and insignificant for analysis i.e. Desc,Emp\_title,application\_type,member\_id

Same Value Columns:

|  |
| --- |
| acc\_now\_delinq |
| chargeoff\_within\_12\_mths |
| delinq\_amnt |
| tax\_liens |
| pymnt\_plan |
| collections\_12\_mths\_ex\_med |
| policy\_code |
| application\_type |
| initial\_list\_status |

Insignificant Columns:

|  |
| --- |
| desc |
| title |
| member\_id |
| url |

* Filling Na values with blanks to maintain uniform in Datatypes.
* Removed suffix with 'xx', ‘months’ in zip code, Term columns to reduce data size and bring consistency.
* Converted Term column into Number format since after removing string this would become left with only integer.

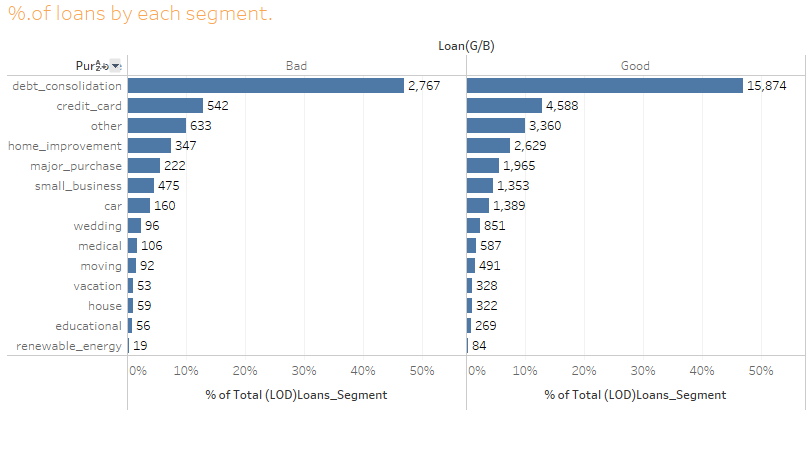
Derived Columns for analysis:

* Adding new column for [pub\_rec\_bankruptcies\_Y/N] to make it values as NO if its Zero.
* Adding Issued Year from issue\_d column to fetch only year from date formatted column.
* For Analyzing Loan Amount whether its Small/Avg/High as per statistical categorial method as first, Avg and Highest Quartiles are finding with variables F\_Q,A\_Q,H\_Q.
* Adding new col as Derived col [typeOfloan ]for interpreting [funded\_amnt\_inv] by above F\_Q,A\_Q,H\_Q values to find outliers.
* Adding Column [Loan(G/B)] to find out if Loan\_status that are "Current", "Issued" and "Fully Paid" can be called “Good Loans” else "Bad Loans".

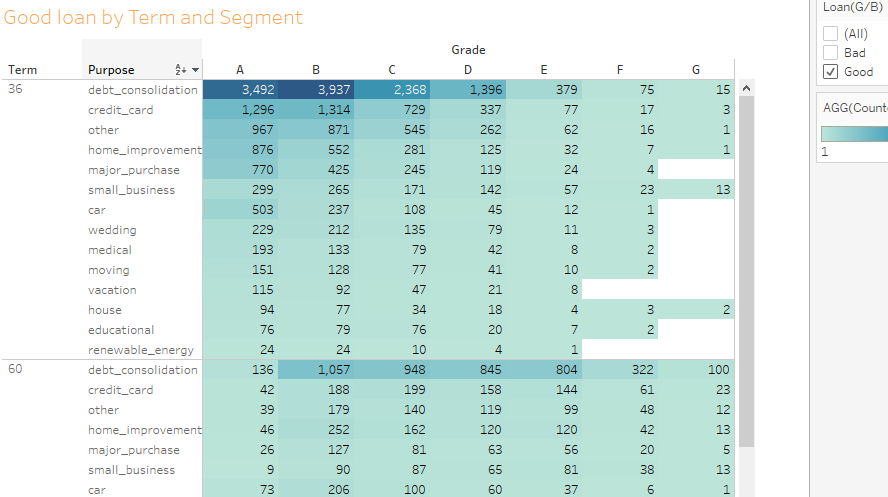
\*Finally Saving Cleaned Data into local drive for further Data visualization and draw inferences out of this data.

**Visualization:**

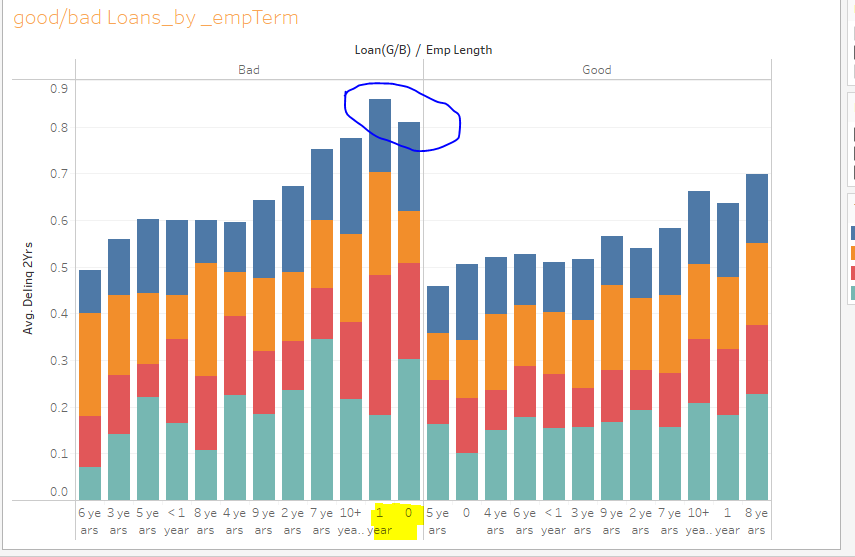
Finding the percentage of Good Loans for each loan segment I have made LOD calculation as SUM([(LOD)Loans\_Segment]) / TOTAL(SUM([(LOD)Loans\_Segment]))., to get accurate calculations regardless of external filters which later if applied external users.



Segmenting the loans by term length (36 or 60 month) and grade by highlight tables for term both 30/60 Months.

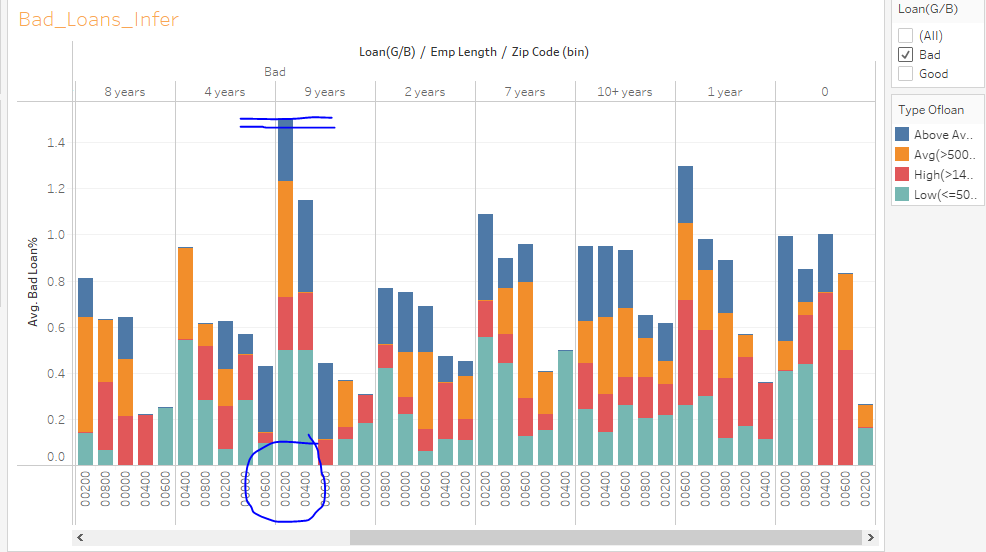


For drawing conclusions on Good loans by which factor influencing the most from given characteristics of the borrowers out of job title and employment & length (years) Emp\_tenure having considerable effect in less than 1 year of experience regardless off title of employment.



Bad loans inferences by Zip can give us in which bins of zip containing arising more bad loans as we see Zip code between

200-400 having more bad loans comparing to other Zip regions.



Alternative Approaches:

We can get more simulated graphs from Python Seaborn and Matplotlib by aggregating custom columns with mean and count values as appropriate to decide Good Vs. Bad loans per each Region, by Borrower Income, Loan\_Status , Grade with subplots by Bar charts.