**Data Quality Insights**

Check the quality of the data in order to verify integrity ad identify bad data issues. Data Quality management is very important data mining step in order to make sure whether the values are present in the attributes that’s required to answer business question.

In order to perform quality check, following steps have been followed:

1. Check incompleteness- check missing values

2. Identify outliers

3. Correctness

4. Bad data

1. **Missing Data**

In data munging part, when multiple data sources are joined together, we might find missing values coming from sources and not in other. Here, after merging loans data with institute data on agency code, respondent id and as of year, following is the count of missing values for each variable for 1.3M records.

1. **Identify outliers**

We would like to find extra spikes in our data. In order to check outliers in loan amount, all the values with fall outside 3 standard deviations are considered outliers.

**3. Correctness**

Loan type description has two values: Conventional and non-conventional while conforming Status can take two values as: conforming and jumbo. Here, following function check whether conforming conventional loans has been correctly tagged under ‘Y’ Flag or not.

**4. Bad Data**

While preforming data quality check, there are two important steps: Identify issues and design strategies to remove issues and clean data. Respondent\_Name\_TS has following data quality issues and this looks like quality issues while data storage.

-Trailing spaces in respondent names: which causing same bank to be represented as multiple.

-Format of Respondent names in inconsistent.

-Some respondent names have FCU in their name whereas same bank appears in dataset again without FCU in its name.

-Spelling mistakes in Names: In order to clean data above issues in variable, routine needs to be developed that verify common data issues such as whitespaces

**Important Variables**

One method to identify important variable is to examine how well each variable explains the variability and its impact on answering key business question. Clearly based on our objective to identify and make strategies for Change Financial in entering home loans market, three important variables which provides directional insights are:

1. Conforming\_Status: In the United States, a conforming loan is a mortgage loan that conforms to GSE (Fannie Mae and Freddie Mac) guidelines. The most well-known guideline is the size of the loan, which as of 2012 based on our data was generally limited to $417 for single family homes in the continental US.

Loans above the conforming limit are known as “jumbo” loans. They are risker to lenders and down payment is usually higher. This information would be a great decisive point for small lenders while expanding their business in home loans market.

2. Loan\_Type\_Description: Currently, conventional mortgages represent around two-thirds of the homeowner's loans issued in the U.S. Some of the important factors to consider from consumer point of view are credit score, size of down payment, etc. So consumer prefer certain type of this loan category. Hence, from financial institution point of view, It would be good to focus not government insured loans for lesser risk involved and security.

3. Loan\_Purpose\_Description: This is the major classification in lending market and given the knowledge based dynamics in each of these categories for Refinance and Purchase, this would help Change Financial to develop different plans instead of one set common rules for both.

Further analysis and data exploration of these three variables can tell a great story about the data.