CAPSTONE PROJECT 1 - DATA WRANGLING

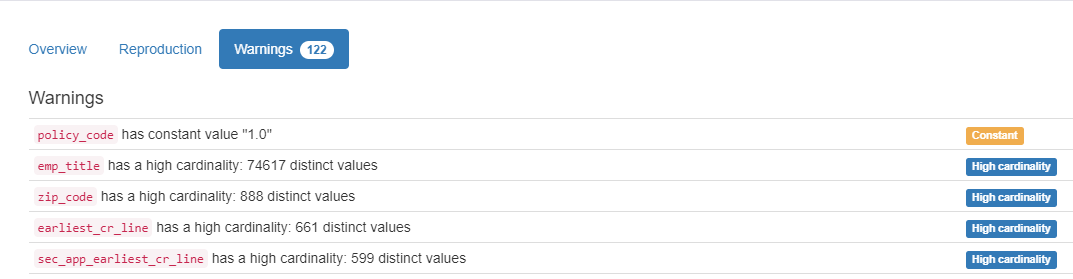
**In my Capstone project I merge 3 files together**:

1. LENDING CLUB accepted\_2007\_to\_2018Q4.csv file that I filter down to application for the 2017-2018 years.
2. THE BUREAU OF LABOR STATISTICS - STATE UNEMPLOYMENT FILE FOR 2017-2018
3. THE WORLD POPULATION REVIEW State Abbreviations File

* From the second file, I removed 'labor force' ,'employment' and ,'unemployment' which I will not use. I also removed the '(R)' from the unemployment variable values and I renamed variables to not include spaces in the names.
* From the third file I removed the 'Abbrev' variable.
* After merging the files together I removed merge key variables the merging process generates. I also removed variables that have 0 non-nulls. Then I removed a variable that has a constant value. I then removed 20 variables that have 100% missing values or virtually 100% (99.9%) as these will have no value to the model.
* I then looked at quite a large number of variables that had missing values where I first took a look at possible outliers to determine what to do before imputing missing values. I ended up finding the outliers and capping the values to 3 standard deviations around the mean for continuous variables. For imputations, I used the mean for continuous variables, -1 for discrete numeric variables, ‘Unknown’ for categorical variables, and ‘Jan-1900’ for a Date variable.



* I then used the Pandas Profiling Report to simplify 2 variables with High Cardinality by keeping the 5 values with the highest frequencies and bunching all others into a new category I named ‘Other’. I dropped another 2 variables with High Cardinality because they were Date variables I will not be using in the model.



* I then removed another variable that was flagged as unique on the Pandas Profiling Report.



* Next I worked through 11 variables that were flagged as Highly Skewed on the Pandas Profiling Report. For those, I identified outliers in the same way as before and I removed them.

