DATA STORYTELLING

After removing the outliers, our next step is to explore the data to understand trends and hidden correlations in the data and visualize them. For this purpose, we will primarily plot graphs using the *'matplotlib'* and *'seaborn'* packages in Python. We will also ask some key questions and try to answer them using Pandas.

Preliminary Questions:

1. First, and probably one of the obvious questions to ask would be ***'How many people have actually even defaulted on their home loans'***? To answer this, we can count the number of rows in the dataset that have a default status of '1' (since 0 indicates no default and 1 indicates otherwise). And we see that ONLY ***9,305*** are defaulters – which is a VERY small percentage of observations/records.

2. Another interesting thing to think about might be ***'How many people have a bad credit-score'***? According to Credit Karma, a score of 700 and above would be considered a good score typically. Using pandas, we can quickly check the count of rows with a credit-score of 700+. We observed that ***~85%*** of the population have a good score – which means that ***15%*** do not, which is the sample of interest to us.

Next step would be to explore the data with visualizations. We will create various plots like Bar plots, Histograms, Scatterplots and Heatmaps to study the data more closely.

- ***Bar plot:*** This is probably one the simplest yet very insightful plots. It gives a quick overview of how categorical variables are spread across the data.

    1. *Channels – Retail, Broker, Correspondent:* A clear trend is that for the least number of loans or few loans (~250,000), home loans originate through Brokers, whereas most happen through Retail channels (1,750,000) followed by Correspondents (~900,000).

    2. *Number of borrowers:* The number of people borrowing a loan is quite evenly distributed between single borrower (1,400,000) and multiple borrowers (1,500,000).

    3. *Loan Purpose:* Indicates whether the mortgage loan is a Cashout Refinance mortgage, No Cash-out Refinance mortgage, or a Purchase mortgage. Majority of them are borrowed for some kind of Purchase (1,400,000). Second in order comes No Cashout

Refiance - loan in which the use of the loan amount is limited to specific purposes. And the least number of loans are for Cashout Refiance - loan in which the use of the loan amount is not limited to specific purposes (only 700,000, which is half of Purchase).

4. *State:* Trend shows that California is the state with the highest number of home loan borrowers, followed next by Texas and Florida.

- ***Correlations:*** A Heatmap is usually used to draw a Pearson-correlation plot of all the numeric variables in the dataset. Using this plot, we observed that numeric variables *CLTV* and *MI%* are correlated and it would be interesting to visualize their relationship.

- ***Scatterplot:*** A joint-plot or a scatterplot can be drawn between the above two numeric features that also shows their distributions on two axes. Using such a plot, we found that a LOT of the observations have their MI% = 0, but their CLTV value uniformly increases from 0 – 100 and beyond. Also, for most observations in the population, the CLTV value is > 60. CLTV refers to the Loan-To-value Ratio and MI% indicates the Mortgage Insurance Percentage.