# **Inferential Statistics Steps – Capstone Project 1**

After completing my Inferential Statistics assignments, I applied this knowledge towards my capstone project 1. Here are some steps that I took to better understand my dataset.

* I created two datasets out of my original datasets. One contained all the loans which were fully paid whereas the second one had all the loans which defaulted.
* I calculated the length of both the datasets to ensure both data frames have more than 30 records so that I can apply Central Limit Theorem on them.
* As both the datasets had more than 30 records, we can apply z-statistics.
* I calculated the mean and standard deviation of these two datasets.
* From the above calculation, I found that the mean and the standard deviations of fully paid and default loans differ slightly. The next step is to determine if this difference is a statistically significant one.
* I created a hypothesis test as stated below:
  + Null Hypothesis: The mean annual income for fully paid and default loans is the same.
  + Alternate Hypothesis: The mean annual income for fully paid and default loans is not the same.

The threshold value of α is assumed to be 0.05. Assuming Null Hypothesis is true

* First, I calculated the difference in mean of both the datasets.
* Then, I calculated the sigma difference.
* Once I had the above two values, I plugged them into my z-statistics formula to calculate the z-score.
* Then, I calculated the p-value to understand the probability of my null hypothesis.
* The p-value turned out to be 0.0 which is less than the accepted threshold of 0.05. Therefore, we reject the null hypothesis. There is a significant difference in the annual income of fully paid and default loans.

For this test, two-sample test was used as we were trying to compare the means of two different groups. Also, the z-statistic was used as the sample sizes for both groups were greater than 30.