**Data Science and Python**

**lessons learned**

**and**

**recommendations**

**Credit One**

1 bigData STreet, Ottawa, ON Canada

**Course 5 Task 2**

**By amarendra ghanekar**

**April 16, 2018**

Table of Contents

[Exploratory Data Analysis 3](#_Toc512028352)

[Lessons learned and recommendations 3](#_Toc512028353)

# Exploratory Data Analysis

For any similar project, it is very critical to have visual output such as scatter plots and histograms as part of EDA. Here are a few key observations from the EDA.

**Age:**

* Age range is 21 to 79 with mean of 35.
* Age can be discretized in 4 buckets

**Limit Balance:**

* Around 66% people have Limit Balance of 200,000
* Third Quartile or Upper quartile (UQ) - 75% people have Limit Balance of 240,000.

**Correlation and Covariance:**

* Use of correlation and covariance also plays a key role in the feature selection.
* Correlation is a measure that determines the degree to which two variables' movements are associated
* Covariance is a measure of how changes in one variable are associated with changes in a second variable. Specifically, covariance measures the degree to which two variables are linearly associated. A positive covariance means that asset returns move together, while a negative covariance means returns move inversely.
* Default payment next moth has positive covariance with PAY\_0 through PAY\_6 and education. Default payment next month has negative covariance with rest of the features. Default payment next month has strong negative covariance with LIMIT\_BAL.

# Lessons learned and recommendations

* Exploratory Data Analysis and data preparation are the most activities in the CRISP-DM model. The more time we spend on these activities, the more we earn about the data and that helps us do better analysis and predictions. Considerable amount of time needs to be invested in this phase of the project in order to arrive at successful predictions.
* Python is fast and powerful programming language but it does have a substantial learning curve.
* It may take considerable time to install Python and related libraries in order to ready for actual development.
* Dora is something worth exploring further for this and future projects.