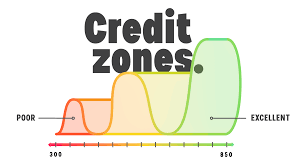
**Credit One**



**Exploratory Data Analysis**

**December 2020**

**Agenda**

1. **Business Value Learnings**
2. **Lessons Learned**
3. **Analysis Recommendations**

**1. Business Value Learnings**

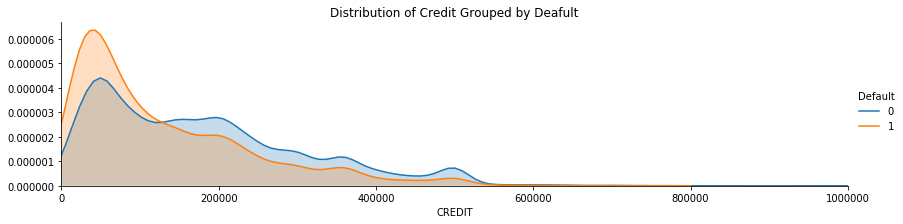
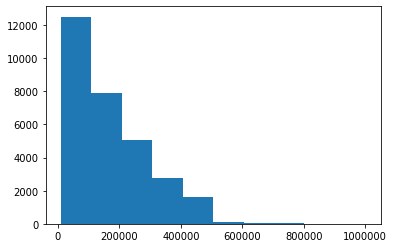
**Correlations and covariance**

Based on the analysis of the data it is observed that there is not a strong correlation between the dependent (Default) and independent variables.  This means that there is no strong relationship, and that in fact variables are hardly related to it.  Nevertheless, collinearity was observed between independent variables.  This is important to identify and take out as it could cause problems when fitting the model and interpreting the results.

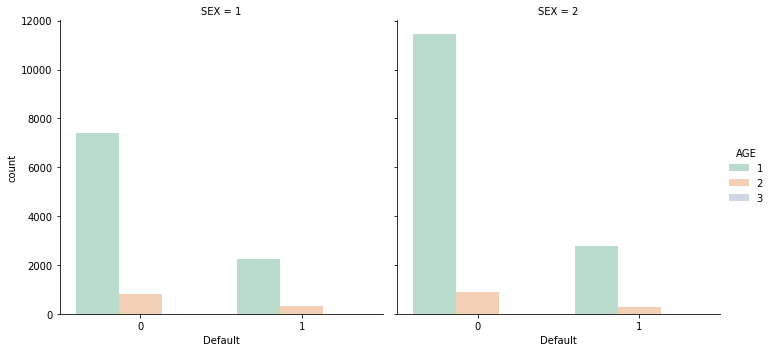
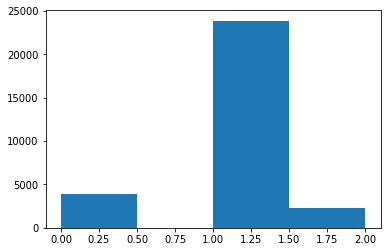
The independent variable has a positive covariance only with the “Education” variable.  Covariance refers to the measure of how two random variables in a data set will change together. A positive covariance means that the two variables at hand are positively related, and they move in the same direction.  It has a negative covariance with all the remaining ones.

**Credit Amounts and Age**

From a credit perspective, it was noted that the majority of the loans given are under $200K.  Following a similar trend, people default more on smaller loans (under $100K) rather than in the larger ones.



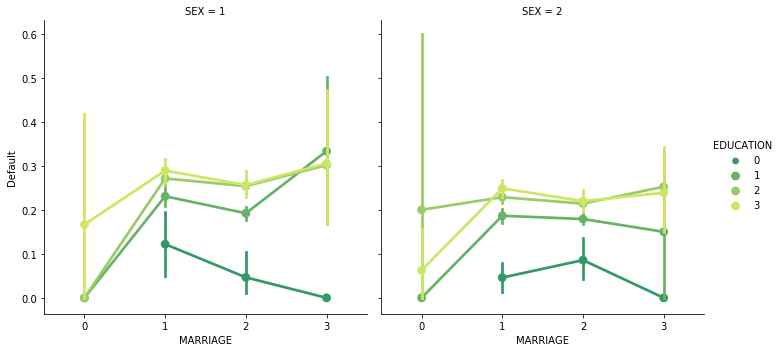
In case of the age variable, most of the loans are given to customers that range from 25 to 50 years old (Bin #1).  It is not a surprise that most of the people that default are under this age range, regardless of sex or other variable.



**Age:** 1. 0 to 25, 2. 25 to 50, 3. 50 to 75, 4. 75+

**Default profile**

Based on the previous analysis, default loans are centered on lower $ amounts and from customers between 25 to 50 years old.  By analyzing the other profile variables, we are able to provide an estimation of what type of customers are more likely to default their loans.



**Default:** 0. No default, 1. Yes Default **/ Sex:** *1. Male, 2. Female* **/ Marriage:** *0. Other, 1. Married, 2. Single, 3: Divorce* **/**

**Education:** *0. Other, 1. Graduate, 2. University, 3. High School*

In the graph the higher the “Default” criteria, the ore likely scenario that a person will not pay their loan

The analysis above shows us that divorced males with “Graduate”, “University”, and “High School” education are the most likely to default their credit.  Important to notice too that single males with the same education are not far behind.

On the other hand, divorced females with “University” and “High School” education more likely to default, following a similar trend to their male counterparts in the next categories.

The above analysis is validated by the table below, which shows how the profiles above carry the most default numbers of the population.



**2. Lessons Learned**

**Learning #1: The power of visualizations**

This exercise required the development of different visualizations, and different slices of the data.  This was crucial in order to validate current observations but also in terms of uncovering potential hidden information.  It also provided easy analysis and digestion of the data to draw conclusions or recommendations.  The important learning is to incorporate a diversity of views and slices of the same data to get a better understanding of it and broaden your perspective.

**Learning #2:  Combine tables with graphs**

As I was going through the data, certain graph visualizations provided percentage patterns that drew certain conclusions.  Nevertheless when you compared the absolute data in tables, you identified that although the graph was accurate from a holistic picture, the conclusions did not really explain the majority of the default cases.  But you could never have deducted that without the comparison of different visualizations.  The learning from this experience is to always compare and contrast, and identify if both options are telling the same story.

**Learning #3: Importance of not having biases**

Based on the exercise done in the previous module, I had already an idea of the correct customer profile that resulted in loan defaults.  And although the data confirmed a lot of my assumptions, there were a lot of other findings I was not expecting, in fact the primary profile identified is not that one I had in mind.  My big learning here is how your biases could derail an objective analysis, and eventually the results of your investigation  Never leave a stone unturned, or a data without analysis.

**3. Analysis Recommendations**

**Customer Profiling**

As we analyze the data, it becomes apparent that being able to profile customers or potential clients becomes an important tool or exercise for the business.  It provides valuable insight to support credit decisions in terms of the amount and type of credit, but also if a person is subject to a loan or not.  This will improve the default indicators the company is facing.

**More requirements for smaller credits**

The data is showing us that most of the defaulted credits are under $100K.  Additionally to the profile exercise, the company could add additional requirements to people requesting this type of loans, like warranties or guarantors.  With this action the company could cover itself, and limit the potential loss.  It could also recover all or part of the defaulted loan.

**Model Definition**

As we reviewed the data we do recommend to treat the creation of the model as a “Classification” exercise. Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category/class to which a new data will fall under. In our case, the variable we would be looking to identify or predict is “Default”. This would provide Credit One with a powerful tool to determine proactively what type of customers are more likely not to pay their loans.