Capstone Project 1: Data Wrangling

Steps Applied to Data Exploration and Manipulation

Downloaded data unto hard drive

Use Jupiter Notebook to Import data

Explore data

Identify data type for each column

Determine the number of unique values per column

View a sample of the data to determine if the data type needs to be changed based on value

Dependents to be converted from object (because of "3+" entry) to integer after simplifying "3+" as "3"

Loan amount is entered in thousands, this was concluded due to the fact that there's no logics in approval for a $98 loan with a term of 360 months and income of $3127, see LoanID LP002502.

Income is entered as monthly (559/614 Applicant Income and 608/614 Co-applicant Incomes are below $10,000)

Loan amount term is in months

Cleaning

Rename columns to remove underscores

The missing values for the LoanAmountTerm column will be replaced by 360, 83% of all applications request a term of 360 and these loans range from $9,000 to $600,000.

The missing values for the SelfEmployed column will be replaced by No, 81% of applicants aren't self-employed.

Drop rows with missing loan amounts and dependents. A summary of the data before and after reveals that there isn't much differences between the means for each numerical variable. Also 94% of the data remained.

Create columns; TotalIncome (Applicant + Co-applicant), MonthlyPaymentNoInterest (LoanAmount * 1000/LoanTerm) and LoantoIncomeRatio (MonthlyPaymentNoInterest/TotalIncome)

The LoanID LP002317 and LP001448 had a similar application with incomes of 81,000 and 23,803. LP002317 was denied a loan therefore 81,000 could possibly be an outlier that can be safety be removed or it could be a yearly income.