10.3 Final Project Step 2

Sharath Chandra Tummanapally

11/12/2021

How to import and clean my data

The datasets I'm using are in csv file format. I am using read_csv to import the data to the dataframe. They are some variables/columns that are not useful for my analysis, they need to be removed.

What does the final data set look like?

```
## 'data.frame':    252000 obs. of  16 variables:
##  $ State            : chr  "Andhra Pradesh" "Andhra Pradesh" "Andhra Pradesh" "Andhra Pradesh" ...
##  $ Bank             : chr  "State Bank of Mysore" "State Bank of Patiala" "Bank of India" "Indian Ba
##  $ Id               : num  82436 159564 129714 242881 213641 ...
##  $ Income           : num  6204574 8786565 5522159 6307868 9840303 ...
##  $ Age              : num  27 42 22 23 24 46 34 34 77 27 ...
##  $ Experience       : num  12 5 6 5 4 10 5 5 19 14 ...
##  $ Married/Single   : chr  "single" "single" "single" "single" ...
##  $ House_Ownership  : chr  "rented" "rented" "rented" "rented" ...
##  $ Car_Ownership    : chr  "no" "no" "no" "no" ...
##  $ Profession       : chr  "Chemical_engineer" "Industrial_Engineer" "Computer_hardware_engineer" "Ma
##  $ CITY             : chr  "Srikakulam" "Hindupur" "Anantapur" "Guntakal" ...
##  $ CURRENT_JOB_YRS  : num  5 4 6 5 4 6 5 5 11 12 ...
##  $ CURRENT_HOUSE_YRS: num  14 12 12 12 11 13 13 13 10 12 ...
##  $ Risk_Flag        : num  0 0 0 0 0 0 0 0 1 0 ...
##  $ InterestRate     : num  8.65 8.3 8.5 8.6 8.65 8.5 8.65 8.6 8.75 8.35 ...
##  $ UnempRate        : num  5.4 5.4 5.4 5.4 5.4 5.4 5.4 5.4 5.4 5.4 ...
```

```
##     State               Bank                 Id             Income
##  Length:252000      Length:252000      Min.   :     1   Min.   :  10310
##  Class :character   Class :character   1st Qu.: 63001   1st Qu.:2503015
##  Mode  :character   Mode  :character   Median :126000   Median :5000694
##                                        Mean   :126000   Mean   :4997117
##                                        3rd Qu.:189000   3rd Qu.:7477502
##                                        Max.   :252000   Max.   :9999938
##
##       Age          Experience    Married/Single     House_Ownership
##  Min.   :21.00   Min.   : 0.00   Length:252000      Length:252000
##  1st Qu.:35.00   1st Qu.: 5.00   Class :character   Class :character
##  Median :50.00   Median :10.00   Mode  :character   Mode  :character
##  Mean   :49.95   Mean   :10.08
##  3rd Qu.:65.00   3rd Qu.:15.00
##  Max.   :79.00   Max.   :20.00
```

```
##
##  Car_Ownership       Profession            CITY           CURRENT_JOB_YRS
##  Length:252000      Length:252000      Length:252000      Min.   : 0.000
##  Class :character   Class :character   Class :character   1st Qu.: 3.000
##  Mode  :character   Mode  :character   Mode  :character   Median : 6.000
##                                                           Mean   : 6.334
##                                                           3rd Qu.: 9.000
##                                                           Max.   :14.000
##
##  CURRENT_HOUSE_YRS   Risk_Flag      InterestRate     UnempRate
##  Min.   :10        Min.   :0.000   Min.   :8.000   Min.   : 1.10
##  1st Qu.:11        1st Qu.:0.000   1st Qu.:8.500   1st Qu.: 4.20
##  Median :12        Median :0.000   Median :8.600   Median : 4.90
##  Mean   :12        Mean   :0.123   Mean   :8.577   Mean   : 7.63
##  3rd Qu.:13        3rd Qu.:0.000   3rd Qu.:8.750   3rd Qu.: 5.60
##  Max.   :14        Max.   :1.000   Max.   :8.800   Max.   :30.70
##                                                    NA's   :2354
```

```
##            State                   Bank     Id  Income Age Experience
## 1 Andhra Pradesh  State Bank of Mysore  82436 6204574  27         12
## 2 Andhra Pradesh State Bank of Patiala 159564 8786565  42          5
## 3 Andhra Pradesh         Bank of India 129714 5522159  22          6
## 4 Andhra Pradesh           Indian Bank 242881 6307868  23          5
## 5 Andhra Pradesh           Andhra Bank 213641 9840303  24          4
## 6 Andhra Pradesh   Union Bank of India 250809 2190403  46         10
##   Married/Single House_Ownership Car_Ownership                Profession
## 1         single          rented            no          Chemical_engineer
## 2         single          rented            no        Industrial_Engineer
## 3         single          rented            no Computer_hardware_engineer
## 4         single          rented            no                 Magistrate
## 5         single           owned            no                    Aviator
## 6         single          rented            no                    Surgeon
##           CITY CURRENT_JOB_YRS CURRENT_HOUSE_YRS Risk_Flag InterestRate
## 1   Srikakulam               5                14         0         8.65
## 2     Hindupur               4                12         0         8.30
## 3    Anantapur               6                12         0         8.50
## 4     Guntakal               5                12         0         8.60
## 5     Gudivada               4                11         0         8.65
## 6 Narasaraopet               6                13         0         8.50
##   UnempRate
## 1       5.4
## 2       5.4
## 3       5.4
## 4       5.4
## 5       5.4
## 6       5.4
```

What information is not self-evident?

Since I'm considering the external factors that could effect laon defaulting. There is no clear correlation between unemployment rate across states in India/Interest rate across banks and loan defaulting/loan risk flag. I need to figure that our during our analysis.

What are different ways you could look at this data?

1. Unemployment Rate against the Loan Risk flag
2. Interest rate against Loan Risk flag

How do you plan to slice and dice the data?

There is a risk_flag variable in our CustLoanInfo_df which I am planning to categorize the customers at high risk of loan defaulting. Firstly, I will use the unemployment rate dataset and check the correlation with the rates based on the location (state) vs the risk flag of the customer from those states in the Customer Loan Data. Secondly, I will use the interest rate across the banks dataset and check the correlation the Lending_bank variable with the risk flag variable of the Customer Loan Data dataset.

How could you summarize your data to answer key questions?

I will use summary() function on the model in providing the R and R-squared if it's a Linear model or Deviations if it's a Logistic regression. It will provide us the information on how predictors influence the outcome.

What types of plots and tables will help you to illustrate the findings to your questions?

I'm going to run Correlation tests, use Scatter plot and histograms.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Given the scenario, I think the research questions could be answered with regression models. I may use machine learning technique, if they aren't answered by outcome of model.