Data Analysis With Python Case Study - 1

Case Scenario:

A consumer finance company specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for a loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
2. If the applicant is not likely to repay the loan, i.e., he/she is likely to default, then approving the loan may lead to a financial loss for the company.

The data given below contains the information about the accepted (Accepted=1) and rejected (Accepted=0) applications.

Data provided is a sample taken from Lending Club:

https://www.kaggle.com/datasets/wordsforthewise/lending-club?select=rejected_2007_to_2018Q4.csv.gz

Variable description:

- Amount.Requested: The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
- Application.Date: The month which the loan was funded
- Loan.Title: The reason for the loan request
- Risk_Score: The lower boundary ranges the borrower's last FICO pulled belongs to.
- Debt.To.Income.Ratio: A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
- Zip.Code: The first 3 numbers of the zip code provided by the borrower in the loan application.
- State
- Employment.Length: Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
- Accepted: For accepted loan, Accepted=1 and rejected loan, Accepted=0.

Problems:

1) Clean the debt-to-income ratio feature by removing any non-numeric characters (like %) and convert it into float type variable.

2) Transform the application date into the "dd-mm-yyyy" format, considering that the current format includes both the date and time components.

3) Extract the month and year from the Date column using appropriate functions.

4) Display the dataframe in decreasing order based on the risk score.

5) How many loan applicants from the state of Texas had their loan applications denied?

6) Remove the records of loan applicants whose debt-to-income ratio fall outside the range of 0 to 99.99.

7) Drop the details of applicants in the finance company whose risk scores are lower than 350 or higher than 850.

8) Estimate the total amount of loan requested by accepted and rejected applicants in the finance company.

9) Calculate the total number of accepted and rejected applicants.

10) What is the average risk score of applicants from the state of California?

11) What is the highest loan amount requested by an applicant who has a job experience of 7 years?

12) Identify the state with the highest number of loan applicants and the corresponding number of applicants.

13) Add a column “Applicant ID” starting from ‘1’ up to the length of the data.

14) Determine the debt-to-income ratio of an applicant with ID 1089.

15) Create “Regions” variable of the applicants based on a lookup table:

lookup_table = {'NJ': 'Northeast','GA': 'Southeast','CA': 'West','OR': 'West','LA': 'South','FL': 'Southeast','AK': 'West',

'CT': 'Northeast','MA': 'Northeast','MS': 'South','TX': 'South','NY': 'Northeast','AZ': 'West','MT': 'West','MD': 'Northeast',

'AR': 'South','MN': 'Midwest','OK': 'South','UT': 'West','NE': 'Midwest','IL': 'Midwest','IN': 'Midwest','VA': 'South',

'MI': 'Midwest','RI': 'Northeast','ND': 'Midwest','HI': 'West','TN': 'South','PA': 'Northeast','CO': 'West','WI': 'Midwest',

'AL': 'South','OH': 'Midwest','NC': 'South','KY': 'South','NH': 'Northeast','MO': 'Midwest','SC': 'South','WA': 'West',

'VT': 'Northeast','DC': 'Northeast','KS': 'Midwest','NV': 'West','DE': 'Northeast','NM': 'West','WV': 'South',

'WY': 'West','ME': 'Northeast','ID': 'West','SD': 'Midwest','IA': 'Midwest'}

16) What is the underlying reason for the borrower with applicant ID 132467 to seek a loan from the finance company?

17) How to identify the data rows that correspond to applicants who have mentioned the term "credit" in the reason for their loan request?

18) Is there a significant difference in the mean loan amounts between high-risk and low-risk loan applicants?

19) Does the risk score significantly impact the loan amount requested by loan applicants?

20) Is there a statistically significant relationship between the risk factor and the likelihood of loan acceptance?

21) Create a visual representation to analyze the following cases

a. Which state had the highest average amount requested in 2008, 2009, and 2010?

b. Which state was most affected by the financial crisis in 2008, based on the highest loan amounts recorded during the years 2008, 2009, and 2010?

22) Analyze the risk score with respect to the employment experience and the loan status of the applicants.

23) Estimate the typical range of debt-to-income ratio of loan applicants with the help of appropriate visualization.

24) How does loan acceptance affect the association between the amount requested and risk score?

25) Identify the year in which the highest loan amount was sanctioned.