*Customer and Loan Data Analysis*

Ronu Skariah   
*Department of Computing*  
National College of IrelandDublin, Ireland  
[x21159840@student.ncirl.ie](mailto:x21159840@student.ncirl.ie)

.

# Related works

Dataset 1 – Customer dataset

This dataset contains basic information of customers. It was created by combining data from two dataset obtained from Kaggle.

There are 13 attributes present in the dataset. Below is the detail of each attribute:

* Cust\_id : Uniques values representing the each unique customers. It is the primary key.
* City : It represents the different city in a state where the customers are from
* State : The place from where the customers are from.
* Cust\_age: Age of the customers
* Date\_created : The record created date.
* Date\_update : The record updated date.
* Email : Email address of the cutomer
* First name : First name of customer.
* Last name : Last name of customer.
* Gender : The Gender of customer
* Income : Income of customer
* Phone\_number : The contact number of customer.
* Postal : Thezipcode of customer.

Research Question

* How are the age and income of the customers related?
* What is the count of male and female customers in a state?
* How is the relation between dependent count, customer age and their income?
* How is the age and education of the customers related?
* What is the marital status of the customer?

# Methodology

## Dataset Acquisition

**Customer Dataset:**

* API: Dataset related to basic customer information is very important in this analysis. To get updated data on regular basis API based system would be suggested and suitable for this. That is the reason we have used <https://www.mockaroo.com/> and generate data to showcase API based system.
* Kaggle: For Analysis purpose Kaggle is the best source to utilize. We got customer data from

Combination two Kaggle datasets :

<https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>

The main objective here is to acquire customer dataset and perform exploratory analysis on the datset.

With the help of this dataset, we can also run various machine learning algorithms but scope of work does not include it. As per the scope of project we have to perform below steps:

* Data Acquisition
* Data Cleaning, Normalization
* Data Visualization
* Store Data in Database.

## Data Cleaning and Transformation

Data collection, cleaning and Transformation has been done through Python using Pyspark, Pandas, Numpy library.

**Customer Dataset:**

Data Preparation – There are various steps involved in it. Below is the detailed description of each step:

* Importing relevant packages like pandas, pyspark, boto3, json, numpy, tkinter to utilize various functions available in it.
* Declare all the variables that are required for further processing.
* Create function to get data from mockaroo api and return PYSPARK data frame
* Create function to read another dataset in json format into pyspark data frame.
* Convert both pyspark data frame into Pandas data frame to perform various transformation and visualization
* Aggregate both Pandas data frame into single data frame
* As per the requirement before performing any data cleaning operation we need to save data into database. We have used AWS S3 bucket to save RAW data.

Boto3 package helps to write the data into respective bucket.

* Now describe and check the shape and other information about pandas data frame created in previous step. With this we will able to understand the shape, min ,max and median value in dataset.
* Next step is to check for missing values in Dataframe by using isnull() function in Python. If there are missing values then we have fill them in by using various methods available in pandas e.g fillna() or bfill

We can plot heatmap to visualize null values

Chart

Description automatically generated

*Fig 1: Null values plot*

* Most of the time in dataset we have “?” as value. Next step is to replace all “?” with NaN value using numpy package.
* Remove irrelevant columns from Dataset. In case of customer dataset, we are getting “\_corrupt\_record” column which we can remove from data frame.
* Check for outliers. If we have outliers in our dataset then we should remove it using the quartile method.

With the help of box plot we can check the outliers in dataset.

Chart

Description automatically generated

Chart

Description automatically generated

*Fig 2: boxplot for outliers*

* Cleaned data is saved into MYSQL database.

# Setting up Pyspark

1. Install Anaconda
2. Install Java 8 and set the path and java home in environment variables
3. Download Spark home using the link [*https://spark.apache.org/downloads.html*](https://spark.apache.org/downloads.html) and select spark version 3.0.0 and Prebuilt Hadoop 2.7 as package type
4. Extract the folder and set the path for Spark home as this path and path as this path up to bin inside the folder in environment variables
5. Download winutils for version 2.7.1 and create a folder winutils and a subfolder inside it as bin and put the file there. Provide the Hadoop home path as this path up to bin.
6. Use pip install pyspark in anaconda prompt.
7. Use pyspark command in anaconda prompt to check if it is successful.

Links for reference : [*https://www.datacamp.com/community/tutorials/installation-of-pyspark*](https://www.datacamp.com/community/tutorials/installation-of-pyspark)

*Data Generation Using Mockroo*

I will create an account in mockroo. Then I will define the schema for customer and provide the function for generation of the data in each fields. Below figure show the overlay of the customer schema in mockroo;

Graphical user interface, application, Word

Description automatically generated

*Fig 3: Schema creation in mockroo.*

Graphical user interface, application

Description automatically generated

*Fig 4: Create API for the schema in mockroo.*

After creation of the schema I will create an api for that schema. API is the acronym for Application Programming Interface, which is a software intermediary that allows two applications to talk to each other. I can use this API to fetch the data from mockroo using python. Once the mockroo set up is done we can create a notebook for ETL process.

*Extraction of Data using Python*

I will start with the first stage of ETL process that includes the extraction of the data from mockroo. I will import the machine learning libraries in my jupter notebook as they will help is handling data more easily. I will import the Pandas, numpy library for the dataframe creation and pre-processing part. Dataframe is a data structure that organize data in a two-dimensional table of rows and columns. The matplot library will be used for creation of visualization using the dataframe.

Graphical user interface, text, application, email

Description automatically generated

I need to communicate with mockroo to get the data from mockroo. This can be done by making a HTTP request using python. The request module will be used for this. I will create a HTTP request to the mockroo api using the get method in request module. The parameters like the api key will be passed along. After successful request to the api it will return the response as json back. Json is a unstructured data we can convert this unstructured data into the structured form and convert it into a dataframe. For conversion of unstructured data to structured data I will use pyspark.

For using the spark functionality I will create a spark session and spark context. They are used to connect with spark cluster, use spark functionalities and create resilient distributed dataset (RDD). Once the unstructured data is structured, I will use pandas to convert it into a dataframe. Below is the sample API response.

Graphical user interface, text, application, email

Description automatically generated

Text, letter

Description automatically generated

*Fig 5: The response JSON from mockroo and converted dataframe using pandas*

1. *Creating Amazon S3 to store the data ‘*

|  |
| --- |
|  |

1.     Sign into to AWS console

2.     Go to S3

3.     From s3 dashboard hit create s3 bucket.

4.     Provide the s3 bucket name automobile-datalake

5.     Select the AWS region as EU Ireland west

6.     Set global in region and region selection is disabled

7.     In advance settings select disable.

8.     Hit the create bucket .

Once the Amazon S3 storage is set I will store the data to the storage. The data is converted to csv and then it is uploaded to the amazon s3. The boto3 client is used to create a connection between aws s3 and python code. For make the connection proper I will create a user in aws IAM roles are generate the accesskey and secret key to authenticate to aws. The following code snippet shows the above illustration.

*Fig 8: Code snippet to upload data to Aws s3*

Another method I have used is to create an interactive dashboard based on the user input. Here the user can select the state from which he wants to retrieve plots and information from the dataset. This functionality is done using a library tkinter. Tkinter library helps to create a UI interface using python code.

Tkinter can be installed in anaconda using the command

* Pip install tk

I will create a pop up box which contains a dropdown of state present in dataset. The user can select the sate from which he wants to retrieve information , including the plots also. This plots include pie chart different customer in the state based on gender, highest of customer, tabular representation customer name and their earnings and bar graph of customer count from the cities in that state. Below code snippet shows the function of the above illustrated.

## Data Visualization

1. Customer Income Analysis:

Plot distribution graph to check income distribution around mean and its standard deviation.

We can plot heatmap for values

Chart

Description automatically generated with low confidence

Chart, box and whisker chart

Description automatically generated

*Fig 7: boxplot and correlation matrix for outliers*

Looking at the distribution plot we can say that there are outliers in the customer dataset

1. Count Analysis:

Chart

Description automatically generated

Chart, histogram

Description automatically generated

*Fig 8: Plot for count analysis*

* There are more male customer than female customers.
* Most of the Customers are from California and Florida.
* All the customers are having an average age between 40-50
* All the customers has greater spending limit and they are almost eligible.
* The Age utilization ratio is high for the customers.

1. Age vs Income

Chart, scatter chart

Description automatically generated

*Fig 9: Plot of Age vs Income*

From the pair-plots , Correlation matrix heat map and scatter plot we can understand that age has a greater influence with earning.

* The middle-aged person earns more percent.
* The Income is less for younger pro-active person
* Old people got overfitted by the others

1. Gender vs Income

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

*Fig 10: Plot of gender vs Income*

* This analysis will show that from the customers the male customers earns more than the female customers.
* The income band is also high for female customers.

1. Customer categories.

Chart, bar chart

Description automatically generated

Chart, pie chart

Description automatically generated

*Fig 11: bat plot and pie plot of gender and income .*

* Men are more attracted to buy new cars rather than women.
* The graph shows double increase in the male than female customer count.
* The below pie chart show the martial status of the customers which show that more than half percentage of customer are married.
* Only 20% are hesitated to provide reveal their relationship.

Chart, pie chart

Description automatically generated

1. Average Customer Age vs Education Level

Chart, line chart

Description automatically generated

*Fig 11: Line graph of age vs education*

* The average customers who have age above 52 had done their postgraduate studies.
* There are less customers in this dataset who has a doctorate.
* Aged customers are well educated.

1. **Loan Dataset**

This dataset contains basic loan information of customers. It was created by combining data from a dataset obtained from Kaggle.

There are 8 attributes present in the dataset. Below is the detail of each attribute:

* Cust\_id : Uniques values representing the each unique customers. It is the primary key.
* Emp\_id : The employment id of the customer.
* Interest rate: The loan amount interest rate.
* Loan\_amount: The loan amount borrowed.
* Loan\_id : unique representation for information It is the primary key.
* Tenture\_yrs: the time for which loan applied
* Vehicle\_id: The unique value for vehicle information
* Loan\_data : Data at which loan was taken.

Research Question:

* How is the distribution of self-employment along with the credit history?
* Does most of the female and male customers have family or not?
* Does customers have more property have completed their loan?

# Methodology

## Dataset Acquisition

**Loan Dataset:**

* API: Dataset related to loan information is very important in this analysis. To get updated data on regular basis API based system would be suggested and suitable for this. That is the reason we have used <https://www.mockaroo.com/> and generate data to showcase API based system.
* Kaggle: For Analysis purposes Kaggle is the best source to utilize. We got customer data from

Combination two Kaggle datasets :

<https://www.kaggle.com/datasets/zhijinzhai/loandata>

The main objective here is to acquire customer dataset and perform exploratory analysis on the datset.

With the help of this dataset, we can also run various machine learning algorithms but scope of work does not include it. As per the scope of project we have to perform below steps:

* Data Acquisition
* Data Cleaning, Normalization
* Data Visualization
* Store Data in Database.

## Data Cleaning and Transformation

Data collection, cleaning and Transformation has been done through Python using Pyspark, Pandas, Numpy library.

**Loan Dataset:**

Data Preparation – There are various steps involved in it. Below is the detailed description of each step:

* Importing relevant packages like pandas, pyspark, boto3, json, numpy, tkinter to utilize various functions available in it.
* Declare all the variables that are required for further processing.
* Create function to get data from mockaroo api and return PYSPARK data frame
* Create function to read another dataset in json format into pyspark data frame.
* Convert both pyspark data frame into Pandas data frame to perform various transformation and visualization
* Aggregate both Pandas data frame into single data frame
* As per the requirement before performing any data cleaning operation we need to save data into database. We have used AWS S3 bucket to save RAW data.

Boto3 package helps to write the data into respective bucket.

* Now describe and check the shape and other information about pandas data frame created in previous step. With this we will able to understand the shape, min ,max and median value in dataset.
* Next step is to check for missing values in Dataframe by using isnull() function in Python. If there are missing values then we have fill them in by using various methods available in pandas e.g fillna() or bfill

We can plot heatmap to visualize null values

Chart

Description automatically generated

*Fig 12: Plot for null values.*

.

* Most of the time in dataset we have “?” as value. Next step is to replace all “?” with NaN value using numpy package.
* Remove irrelevant columns from Dataset. In case of customer dataset, we are getting “\_corrupt\_record” column which we can remove from data frame.
* Check for outliers. We will plot the violin graph and distribution plot to check for the outliers The below figures represent these plots

Chart, radar chart

Description automatically generated

*Fig 13: Plot for Outlier detection*

* Looking at the plots it is clear that the distribution for application income, loan amount, co-applicant income has positive skewness and hence they have outliers.
* Cleaned data is saved into MYSQL database.

## Data Visualization

Loan Amount Analysis:

Plot distribution graph to check income distribution around mean and its standard deviation.

We can plot heatmap for null values

Chart, treemap chart

Description automatically generated

*Fig 14: Correlation matrix plot.*

Looking at the distribution plot we can say that there is a positive correlation between the loan amount and the income of customer.

Missing value Analysis:

Chart, bar chart

Description automatically generated

*Fig 15: Bar graph for missing value.*

* The male applicant is higher than the female applicant and there are missing value
* Married applicants are higher than unmarried so the propability for co-applicant be their wife will be more.
* More than 60% of the applicants are educated.
* 80% of the applicant are not seld employed.

Gender vs Married

Chart, bar chart

Description automatically generated

*Fig 16: Plot for Gender vs Martial status.*

* The graph show that most the male applicants are married compared to female applicants
* The unmarried male applicants count is also higher than female.

Self-Employment vs Credit History

Chart, bar chart

Description automatically generated

*Fig 17: Plot for self-employment vs credit history*

* Non-self-employed applicants have great credit history rather than self-employed applicants.

Property Area vs loan Status

Chart, bar chart

Description automatically generated

*Fig 18: Plot for property area vs loan status.*

* Most of the accepted loan has a property in the semi-urban area rather than a Rural or Urban

Customer and Loan data combined

Chart, bar chart

Description automatically generated

*Fig 19: Plot for Gender vs Loan Status*

*Diagram

Description automatically generated*

*Fig 17: Plot for total loan applied and not applied customer count.*

Creating a MySQL of Azure virtual machine.

1. Create a azure account using an emailed
2. Go to virtual machine and click on create and from option select virtual machine from predefined image.
3. Provide name and size of virtual machine, windows 11 pro as OS and a resource group.
4. Continue to last step by clicking next.
5. Click review and create .
6. Once the vm is up go to networking in overview and provide the inbound and outbound port rule for the port 3306.
7. Connect remotely to the vm.
8. Install MySQL inside the vm.
9. Check the connection by using a SQL workbench.

# Summary

The primary goal of this study is the creation of interactive dashboard that can visualize all the information the particular user (Automobile dealers) wants for all the automobile booking. This project consists of the following stages:

* + - Data Etraction
    - Data cleaning
    - Data Storage
    - Data Visualization

I was able to combine get the unstructured data using api call from python and convert it into a structured data using spark and pandas and combined it with the data from another source. Once the data is extracted I was able to store it in NoSQL database and perform data cleaning and preprocessing and store it back into MYSQL database which was installed in a VM and created that VM in Azure. Also I have performed the visualization which gave the solution for the research question

* + The average income is for middle-aged customers
  + Male customers have bought more cars than female customers
  + The aged people more dependent count and their income are less.
  + There is a percentage of aged people who have more education among the customers.
  + The customers who are not self-employed has good credits than other.
  + Male customers have family compared to the female customers.
  + The customer with semi urban area hasn’t completed their loan.

In this project I used:

Databases:

a) Azure SQL VM relational database

b) Amazon S3 (except SQL)

Programming language: Python

# References

1. Peter E Hickman, Gus Koerbin, Julia M Potter, Nicholas Glasgow, Juleen A Cavanaugh, Walter P Abhayaratna, Nic P West, Paul Glasziou, Choice of Statistical Tools for Outlier Removal Causes Substantial Changes in Analyte Reference Intervals in Healthy Populations, Clinical Chemistry, Volume 66, Issue 12, December 2020, Pages 1558–1561, <https://doi.org/10.1093/clinchem/hvaa208>
2. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May;64(5):402-6. doi:10.4097/kjae.2013.64.5.402. Epub 2013 May 24. PMID: 23741561; PMCID: PMC3668100.
3. MILOVANOVIĆ, I. (2013). *Python data visualization cookbook*.