ISM 6208 – Data Warehousing

Summer 2022

Final Project

Group 11
Shabana Ajamal Hannure
Syed Omar Farooq Ali
Kaushik Vaka
Sai Sujitha Reddy Mullangi
Srinivasa Madhav Amrut Varma Vegesna

Contents

1: Executive Summary

Most of the financial institutions are running Their operations smoothly and profitable way without any Interruptions with the help of data analytical techniques. This study will be able to enhance the business's ability to Expand its market by providing meaningful and key analysis of consumer behavior. Financial institutions should have Proper parameters to identify the right customer base with the capacity of their repayments. To identify those Parameters, BI technologies, and the data warehouse Techniques such as inspecting, cleansing, transforming, and Modeling were used to convert data to meaningful Information. The star schema is used for this data warehouse Design which includes one fact table surrounded by several Dimensions. Those factors will be evaluated by using A decision tree in future works. This will increase the loan Collection efficiency.

Introduction:

Borrowing credit or loans from financial institutions by individuals, small organizations, or large organizations invest in self-employment. Individuals who borrow a certain portion of money at one time repay that with certain steps within the realistic short or long time period. Most of the individuals borrow for educational purposes or professional reasons with their repayment capacity calculating by themself with their monthly wages. Certain organizations in the country borrow the amount from the institution to invest in certain projects in which they are willing to expand their business or gain more profit from that. They are planning to repay that amount with the profit and income which they will achieve by the project they invest in over the time period. Financial institutes such as banks, leasing companies, and other credit departments are willing to earn their profits by funding or giving loans or credits to such organizations or individuals by taking risk of repayment over the period they agree to. This profit-earning cycle has risk factors and those factors should be calculated by the institution before they release the amount to borrowers.

2: Problem Statement

The company wants to automate the loan eligibility process (real time) based on customer details provided while filling in an online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customer segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a partial data set.

When the borrower applies for the loan or credit from a certain finance institute, whether the loan or credit is accepted or rejected according to the screening criteria. After a certain time period, the accepted borrower receives his/her loan. This can be analyzed further with the factors which we obtain from a decision tree. To accomplish the analysis by a decision tree, sample data was collected by certain financial institutions. To discover them, Microsoft BI tools were used to do certain operations such as inspecting, cleansing, transforming, and modeling to convert useful information.

3: Literature Review

We have read multiple research papers regarding anti money laundering and bank loan problems.

[1] AbrehamGebeyehu (2002). “Loan repayment and its Determinants in Small-Scale Enterprises Financing in Ethiopia: A Case of Private Borrowers Around Zeway Area”, M. Sc. Thesis, Addis Abeba University.

[2] Ted E. Senator, Henry G. Goldberg, Jerry Wooton, etc., The financial crimes enforcement network AI system (FAIS) identifying potential money laundering from reports of large cash transactions[J], AI Magazine, Vol.16, No.4, pp. 21-39, Winter 1995.

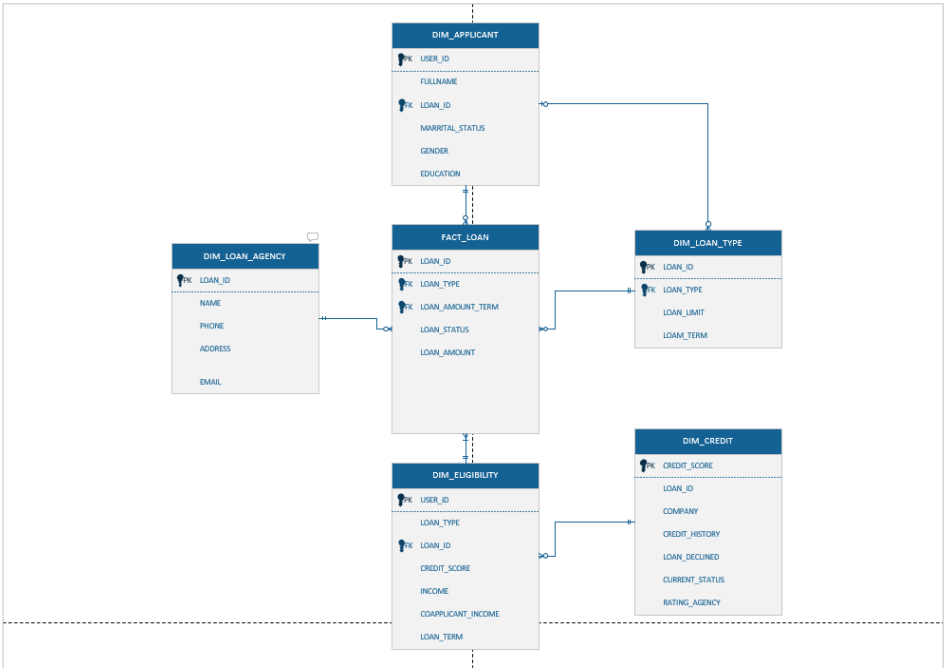
[3] Safavin,S.R., Landgrebe,D. A survey of decision tree classifier methodology [J]. IEEE Transactions on Systems, Man and Cybernetics, Vol.21, No.3, pp.660- 667, April 1991

4: Data Collection and Preparation

The sample data set was collected for the analysis from the Kaggle.com website. Among the larger scale of data set factors were identified to analyze, such as profession, sex, age range, locations wise. The borrower’s repayment amounts, arrears amount, arrears days for the month or quarter, were derived from that source of data set. Before loading data to the data warehouse, there was a workaround to accomplish the task such as cleansing data, conversion, filtering as per needs, sorting, joining, aggregating, and lookup.

5: Database Design

Star schema is used for this study data warehouse design which includes one fact table surrounded by several dimensions. Every fact points to one tuple of each dimension with additional attributes. Also, that does not capture hierarchies. In the star schema, dimension tables will not join each other, and every dimension table joins with the special key called surrogate keys (SKs) with the fact table. Surrogate keys are normally integer values, which is a unique value assigned to each row. SKs will play a very important role in the data warehouse, it helps to protect the data warehouse from unexpected administrative changes, and it helps to updates and inserts as well as tracking of the changes in the dimensions.

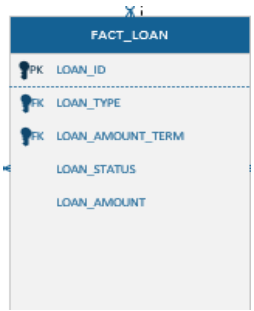


5.1: Transactional Models

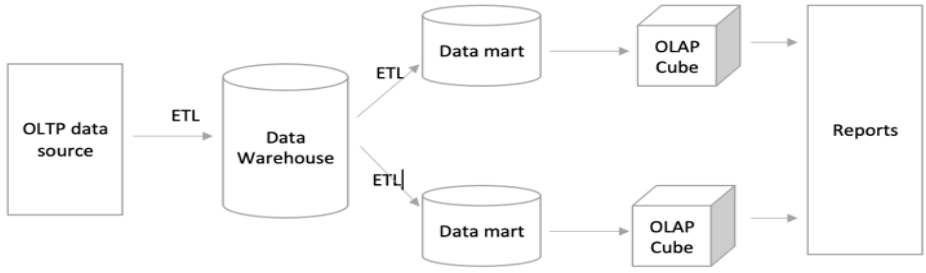
The data warehouse was designed and referred to by this study which has one Fact table called "FACT_LOAN" with five-dimension tables.

5.2: Dimensional Models

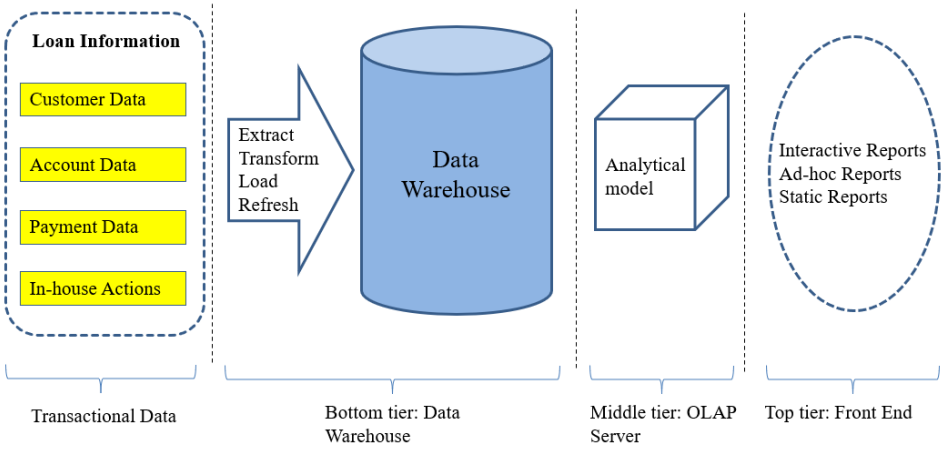
This data warehouse consists of dimensions DIM_LOAN_TYPE, DIM_APPLICANT, DIM_ELIGIBILITY, DIM_LOAN_AGENCY, AND DIM_CREDIT. Most of the data warehouse uses date dimension for the performance reasons and the functional reasons, that consist of the date hierarchy and several other attributes. This study contains two hierarchies for that with quarter and month name as well as numbers for the user's perspective. The focus of this analytical system is to check the actions of the borrowers and due to that FACT_LOAN fact table consists of type, term, amount & status.



When considering this approach, it is like the Inmon approach which is known as “Father of data warehousing”.



This framework also consists of OLTP data source that loads data to the warehouse which is designed via ETL technology and process cube to browse data as multidimensional with several technologies that enable slicing, dicing, roll-up, drill-down, and pivot.



ETL (Extract, Transform, Load)

ETL is a process of loading data from source to destination. ETL defines Extract, Transform, and Load. Extract in the sense extracting data from the source. That source may not be an SQL environment that could be several types of sources. In this operation, we will face extracting data from many systems platforms such as flat files, web servers, emails, images. Due to this we are extracting several types of data into one database called staging area. You must do cleaning and changes to the operational data that simplifies the building summaries that help to avoid slowness, security matters, impossibility, also it avoids the conflicts between source systems. That can be used as a backup option as well.

The second stage of the ETL is transform, it consists of several rules before loading data to the destination. With this stage filtering, sorting, pivoting, validating, deriving columns, joining or merging, splitting, lookups operations can be performed. The above rules vary between the business needs. The approach mostly uses derive functions in this stage to derive and convert data to certain types.

The third and final stage is loading. This stage will check the error logs and send a notification upon that to the administration or the users regarding the availability of the data that can be performed after loading data into the destination.

6: Exploratory Data Analysis (EDA)

Data Analysis is a process to discover useful information for business decision making. This is used to extract useful information from the raw data to make a day-to-day decision based on that information. That may vary according to business needs or individual needs. Analysis is most of the time based on the future prediction by looking into past information or patterns that we have discovered, something like forecasting weather and environmental factors.

There are several types of data analysis techniques such as Text analysis, statistical analysis, predictive analysis, diagnostic analysis, and prescriptive analysis.

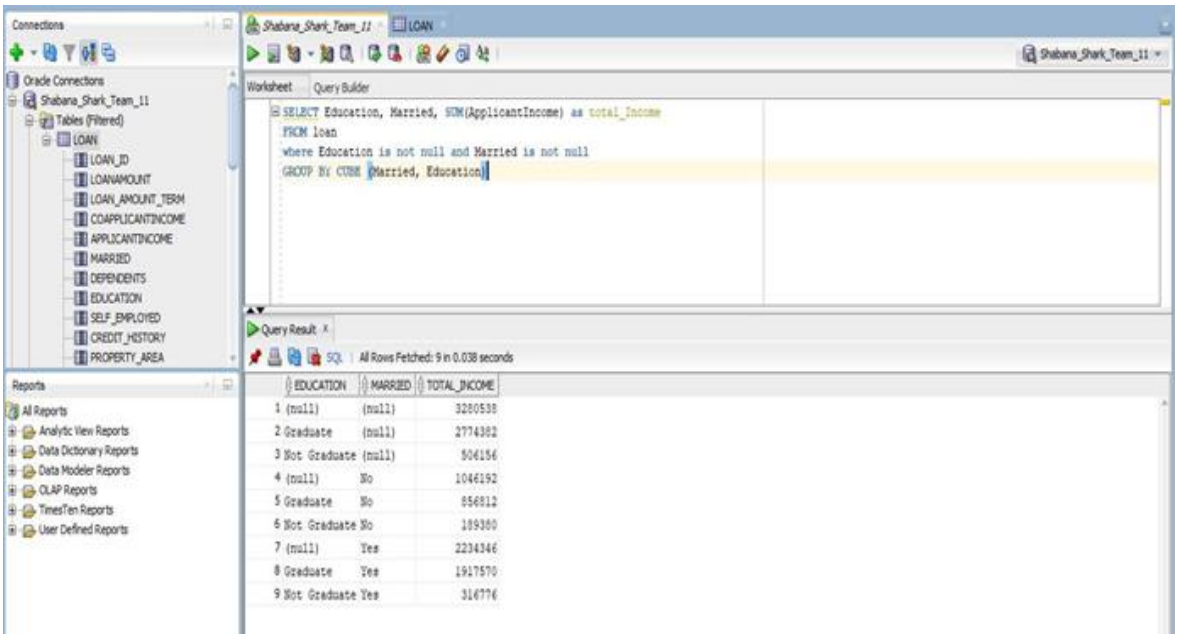
There are several tools to accomplish the data analysis task. This project has used oracle SQL for writing analytic queries and we used power BI and Tableau to deliver our framework for the users.

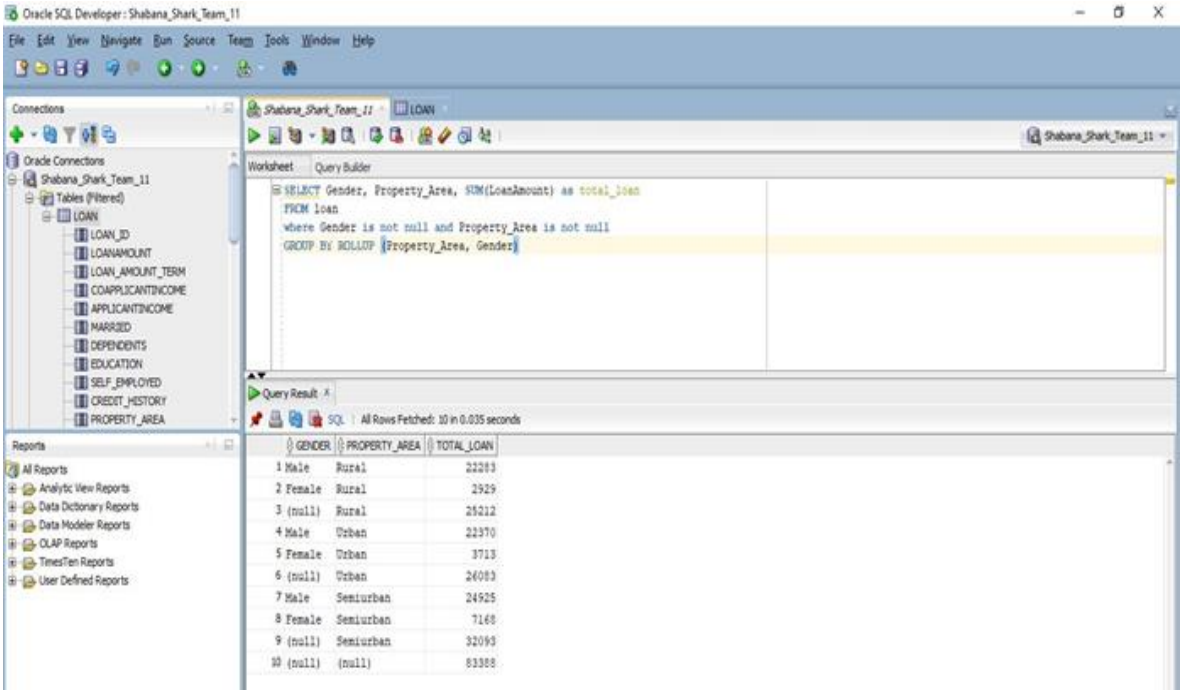
7: Reporting, Modeling and Storytelling

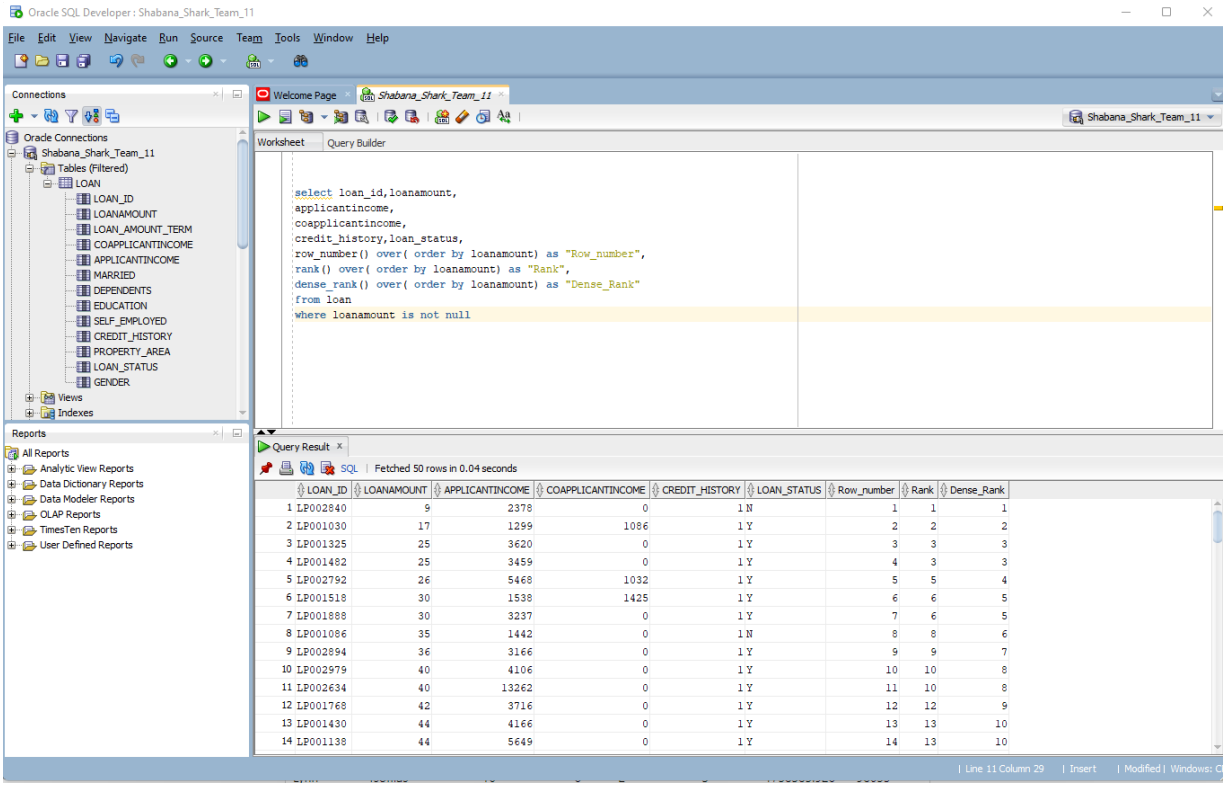
The reporting, modeling and storytelling section is the focus of the project. You are free to pursue one or more of these activities: reporting relies heavily on analytic SQL to generate insightful tables or crosstabs, modeling draws on machine learning algorithms for classification and prediction, while storytelling uses data visualization to develop a narrative. Your project should have a definite focus on modelling or storytelling, but aspects of each of these activities can be used.

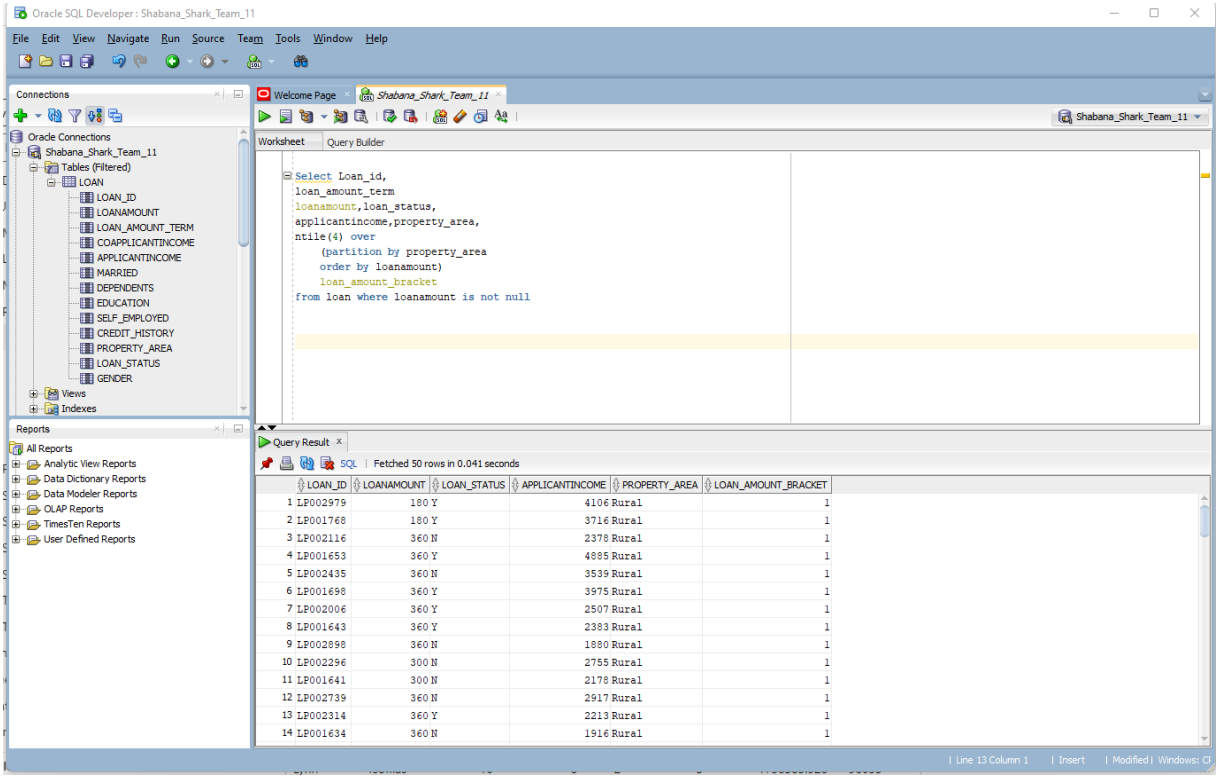
Reporting means illustrating information in a proper understandable manner to users. There are several technical tools such as excel, PowerBI reporting, SQL Server Reporting Service, Tableau and so on. In this approach, we have used Tableau features to represent the dashboard with the visualizations. In order to illustrate the data, this project has utilized Tableau visualizations and the dashboard for management using graphs.

Analytic Queries



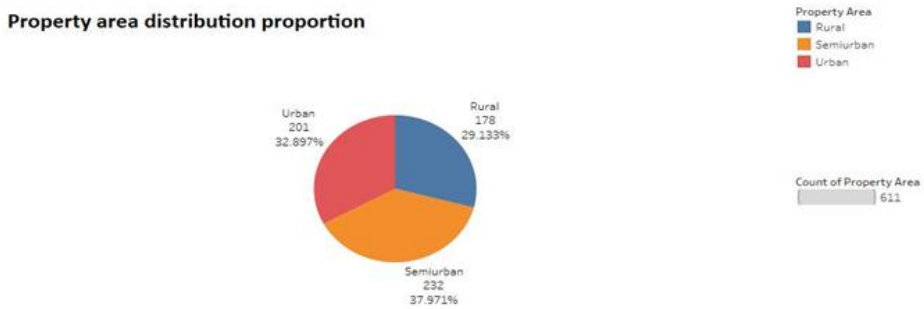


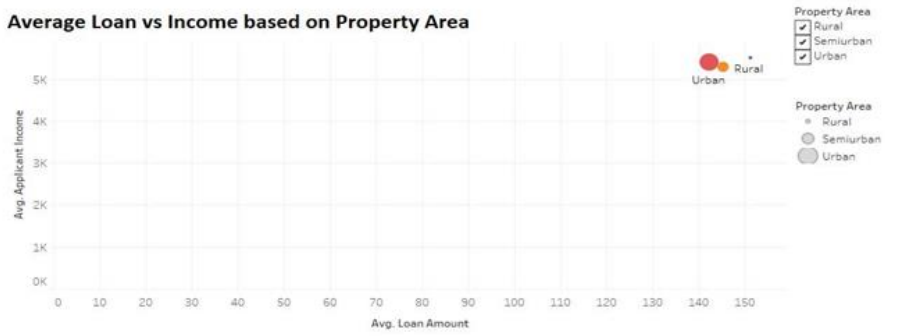

Oracle SQL Developer : Shabana_Shark_Team_11
File Edit View Navigate Run Source Team Tools Window Help
Connections
Oracle Connections
Shabana_Shark_Team_11
Tables (Filtered)
LOAN
LOAN_ID
LOANAMOUNT
LOAN_AMOUNT_TERM
COAPPLICANTINCOME
APPLICANTINCOME
MARRIED
DEPENDENTS
EDUCATION
SELF_EMPLOYED
CREDIT_HISTORY
PROPERTY_AREA
LOAN_STATUS
GENDER
Views
Indexes
Reports
All Reports
Analytic View Reports
Data Dictionary Reports
Data Modeler Reports
OLAP Reports
TimesTen Reports
User Defined Reports
Welcome Page
Shabana_Shark_Team_11
Worksheet
Query Builder
select loan_id,loanamount,
applicantincome,
coapplicantincome,
credit_history,loan_status,
row_number() over(order by loanamount) as "Row_number",
rank() over(order by loanamount) as "Rank",
dense_rank() over(order by loanamount) as "Dense_Rank"
from loan
where loanamount is not null
Query Result
SQL | Fetched 50 rows in 0.04 seconds
LOAN_ID LOANAMOUNT APPLICANTINCOME COAPPLICANTINCOME CREDIT_HISTORY LOAN_STATUS Row_number Rank Dense_Rank
1 LP002840 9 2378 0 1 N 1 1 1
2 LP001030 17 1299 1086 1 Y 2 2 2
3 LP001325 25 3620 0 1 Y 3 3 3
4 LP001482 25 3459 0 1 Y 4 3 3
5 LP002792 26 5468 1032 1 Y 5 5 4
6 LP001518 30 1538 1425 1 Y 6 6 5
7 LP001888 30 3237 0 1 Y 7 6 5
8 LP001086 35 1442 0 1 N 8 8 6
9 LP002894 36 3166 0 1 Y 9 9 7
10 LP002979 40 4106 0 1 Y 10 10 8
11 LP002634 40 13262 0 1 Y 11 10 8
12 LP001768 42 3716 0 1 Y 12 12 9
13 LP001430 44 4166 0 1 Y 13 13 10
14 LP001138 44 5649 0 1 Y 14 13 10
Line 11 Column 29 | Insert | Modified | Windows:


Oracle SQL Developer : Shabana_Shark_Team_11
File Edit View Navigate Run Source Team Tools Window Help
Connections
Oracle Connections
Shabana_Shark_Team_11
Tables (Filtered)
LOAN
LOAN_ID
LOANAMOUNT
LOAN_AMOUNT_TERM
COAPPLICANTINCOME
APPLICANTINCOME
MARRIED
DEPENDENTS
EDUCATION
SELF_EMPLOYED
CREDIT_HISTORY
PROPERTY_AREA
LOAN_STATUS
GENDER
Views
Indexes
Reports
All Reports
Analytic View Reports
Data Dictionary Reports
Data Modeler Reports
OLAP Reports
TimesTen Reports
User Defined Reports
Welcome Page
Shabana_Shark_Team_11
Worksheet
Query Builder
Select Loan_id,
loan_amount_term
loanamount,loan_status,
applicantincome,property_area,
ntile(4) over
(partition by property_area
order by loanamount)
loan_amount_bracket
from loan where loanamount is not null
Query Result
SQL | Fetched 50 rows in 0.041 seconds
LOAN_ID LOANAMOUNT LOAN_STATUS APPLICANTINCOME PROPERTY_AREA LOAN_AMOUNT_BRACKET
1 LP002979 180 Y 4106 Rural 1
2 LP001768 180 Y 3716 Rural 1
3 LP002116 360 N 2378 Rural 1
4 LP001653 360 Y 4885 Rural 1
5 LP002435 360 N 3539 Rural 1
6 LP001698 360 Y 3975 Rural 1
7 LP002006 360 Y 2507 Rural 1
8 LP001643 360 Y 2383 Rural 1
9 LP002898 360 N 1880 Rural 1
10 LP002296 300 N 2755 Rural 1
11 LP001641 300 N 2178 Rural 1
12 LP002739 360 N 2917 Rural 1
13 LP002314 360 Y 2213 Rural 1
14 LP001634 360 N 1916 Rural 1
Line 13 Column 1 | Insert | Modified | Windows:

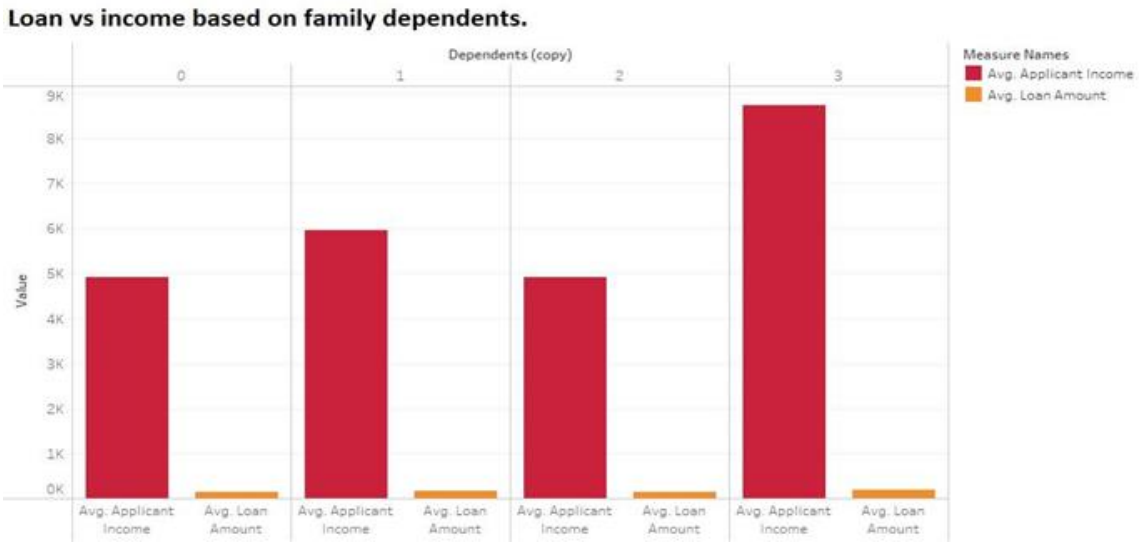
Data Visualization:

From the above graph it's clear that for females there is not much difference in their average loans or income whether they're graduate or not. A female without a degree earns almost the same as one with a degree on average annually. However, with men we see a slight difference. On average men with graduate degree earn much more than men without the graduate education annually. Now comparing the 2 genders, men are slightly more in debt than women in general. Similar is the case with income for the graduates. In contrast for non-graduates on average women earn slightly more than men annually.





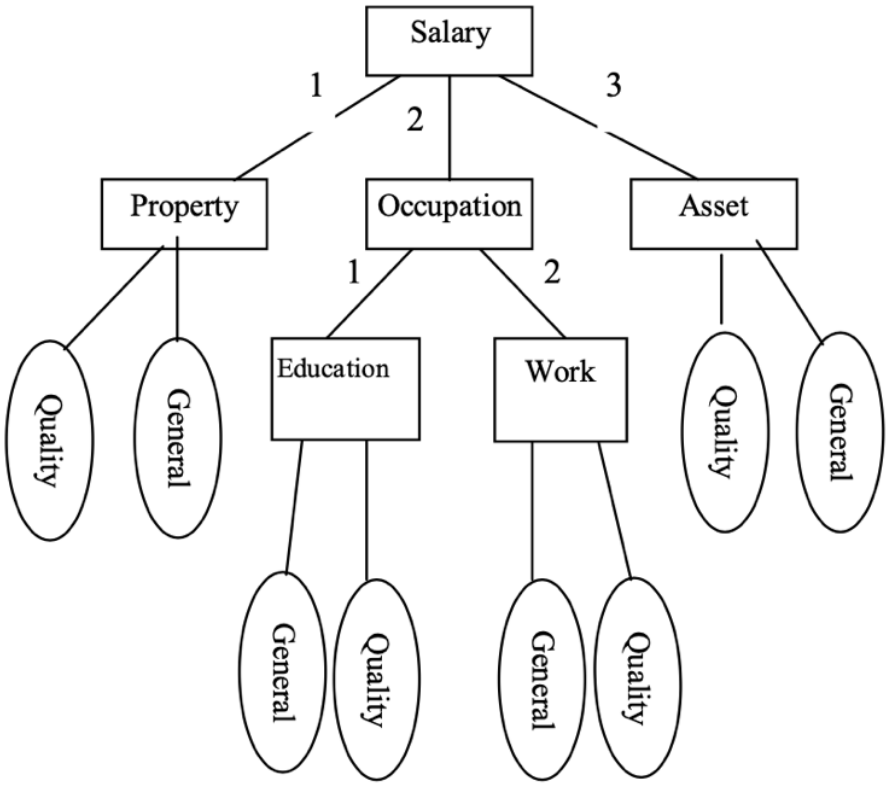
The above dashboard represents a pie chart with the distribution proportion of the type of property areas people live in. So, a greater number of people live in sub-urban areas with about 40%. The rest is almost divided equally into rural and urban areas with over 30% of total population each living there respectively. Below is also a scatter plot of average annual income vs loan amount of people living by the particular property area types. In general, there is not much difference between either the income or loan amount for people based on their living area type.



The above graph shows us the income vs loan amount for people based on their dependent family size. It's evident that the ones with highest dependents earn more annually on average i.e., about $8700 while the rest are about 4000-6000$. Whereas the loan amount is almost the same for people with any dependent family size with people of 3 dependents being the highest for 190 and people with 0 dependents the lowest for about 130.

7.1: Feature Selection

This project has used a decision tree to identify the categories which will affect the repayment, arrears amount, and the response of the borrowers using the data which gives output from the above tools. The decision tree starts with the root node which represents the sample data set. The second step or the second level is called a leaf, this level illuminates measure and the condition of the decision tree second level. All the attributes and discrete values are classified. The attributes of the values must be discrete. In the third step branch and the sample, values will be divided into other branches. If the branch has no sample, that type contains most of the sample that was created as a leaf. That defines the basic decision tree.



In this project, we have used logical regression to predict the category of the loan borrowers that find the best fitting model for the individual case's relationship between response and explanatory variables. The Bayesian logistic regression method is used to predict the repayment efficiency of the borrowers. As a problem, if there is high repayment efficiency, the relationship between the customer and the credit institution is high due to the help of the next higher amount obtained from the credit loan department or the financial institution. To change this, we considered some socio-economic factors. The main factors from the institution side are tight control, loan officer intensive, loan collection, the interest rate charged affect the repayment rate. The socio-economic factors from the customer side are marital state, gender, education, and income level. This project has answered the following questions regarding the repayment factors at the end of the project.

1. What are the main socio-economic factors?
2. What are the business and loan-related factors?
3. What are the major challenges faced by customers and the institute?

The data collection method was a structured questionnaire. We have distributed the questionnaire among the people with the respective population. This includes social attributes, household characteristics, income, assets, financial characteristics such as credit and savings. Logistic regression analysis is an extended technique of multiple regression analysis which is used to identify the categorical variables. We have considered the ratio of the probability of success, and this is the logistic model we used,

$$\frac{P(x_i)}{1-P(x_i)} = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}), i = 1,2,\cdots n$$

We have used Bayesian Logistic Regression to come up with the result. As a result, we came up with, out of 340 borrowers, 38.53% are efficient on repayment and 61.47% not efficient at the time of data collection on them study 11.8% for agricultural, 22.6% trades, 21% small enterprises, 10.9% general loans, 8.8% handcraft. With regards to the sex, they found out 38.2% female and 61.8% male borrowers.

7.2: Discussion

Very important factors include the data analysis, that contains the non-functional requirements, which most of the parties do not consider and may not be known. This can be defined as quality attributes, such as usability, availability, security, and reliability. Those important factors must be involved with the data. If we are working with banking data sources those should be considered mainly due to security. Availability which means that data should be available any time anywhere to access for the users with the credentials they provide, credential? Yes, when we get with the credentials The first thing that comes up is security. The majority nowadays most organizations consider those factors. Analysis systems with nonfunctional requirements provide a better framework.

8: Conclusions

In conclusion, some financial institutions are facing a huge risk due to the failure of recovering loans and collections on time. To avoid this and make decisions on past data by analyzing them. This project was implemented in a framework. It enables you to fetch the reports as per the user requirement with the available data set. For this, Tableau tools were used in order to develop ETL and Cube. This project was conducted to identify the borrower's response. As a result, the researcher was able to identify different key response times of loan borrowers based on several parameters such as age groups and gender. For future work, we have decided to implement a decision tree-based algorithm with classification and regression techniques.

9: References

1. SU-NAN WANG, JIAN-GANG YANG. A MONEY LAUNDERING RISK EVALUATION METHOD BASED ON DECISION TREE, College of Computer Science and Engineering, Zhejiang University, Hangzhou 310027, China 2Shanghai Pudong Development Bank, Shanghai 200002, China 2007.
2. JI Chengjun, WU Lijun, LI Jinping. The Application of the Decision Tree Analyzes in the Credit Card, Department of Management, Liaoning Technical University, Huludao, China
3. Yonas Shuke Kitawa, Nigatu Degu Terye. (2020/10/30). Statistical Analysis on the Loan Repayment Efficiency and it's impact on the, Borrowers: A case study of Hawassa city, Ethiopia Available :http://article.sciencepublishinggroup.com/html/10.11648.j.ajtas.20150406.28.html
4. (202/09/13), Decision tree algorithm example in data mining. Available :https://www.softwaretestinghelp.com/decision-treealgorithm- examples-data-mining/
5. Wang, J. Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications. Hershey: Information Science Reference, 2008.
6. Microsoft (2020/07/22). Analysis and reporting with Microsoft business intelligence (BI) tools. Available : https://docs.microsoft.com/en-us/sql/reportingservices/ choosing-microsoft-business-intelligence-bitools- for-analysis-and-reporting?view=sql-server-ver15\
7. Create an Extended Date Dimension for a SQL Server Data Warehouse, Available : https://www.mssqltips.com/sqlservertip/5553/create-anextended- date-dimension-for-a-sql-server-datawarehouse/.
8. Dinesh Asanka, MSSQLTips.com., What is Analysis Services?, https://docs.microsoftus/analysisservices/