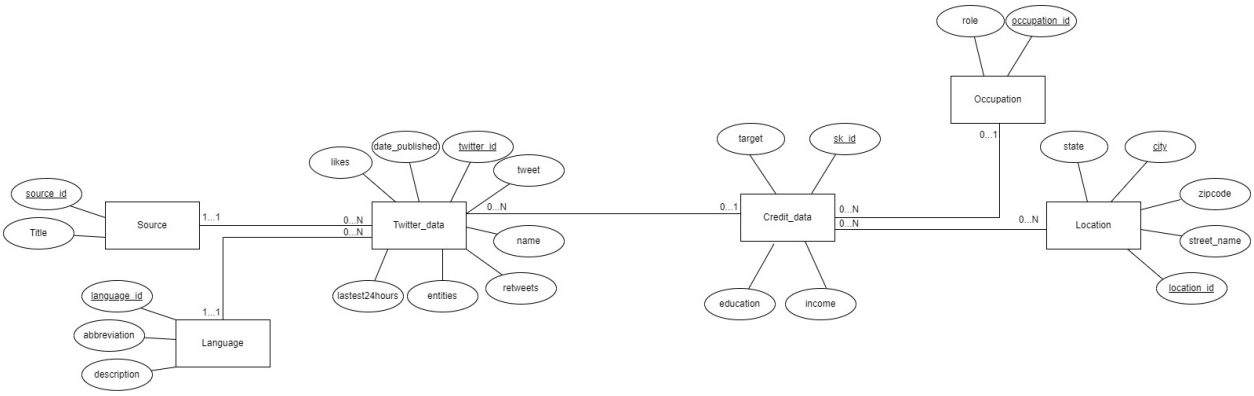
Ram Vaghani (NUID: 002704237)

DMDD ASSIGNMENT 3
Credit Loan Analysis Readme file

Description:
The dataset aims to predict whether the client will be able to pay the loan after some X days based on some features like they employed, if they own home, occupation they belong to, their annual income, etc (more details are given in the table below).

Data Sources:
For this analysis, we have scrapped a twitter data. In this twitter data was formatted where the **language_id** from the table **language** and **source_id** from the table **source** act as **foreign key** to the main table **twitter_data**.
In the other table, credit loan data, customer's details asking for the loan is mentioned. This table to connected to the properties they own and the property location detail. These customers could have also tweeted regarding the credit loan. These customers are working in some occupation, the **occupation_id** from the **occupation** table act as the **foreign key** to main **credit** table.

ERD:
The below ERD shows how all the entities relate to each other (Also feedback-rectify from Assignment-2)

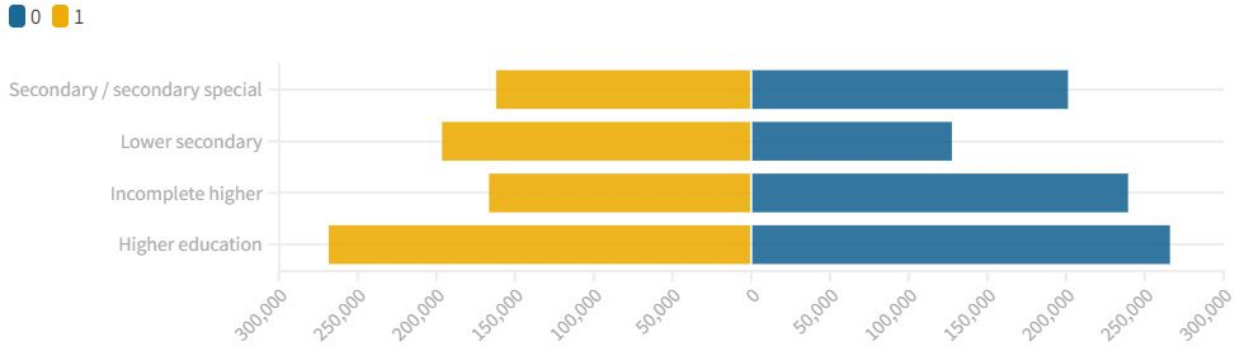


Data Accuracy/Validity:
Based on entities mapping the ERD all the validations have been validated while creating all the tables. Since the customers are only from USA their property evaluation is also supposed to be from USA.

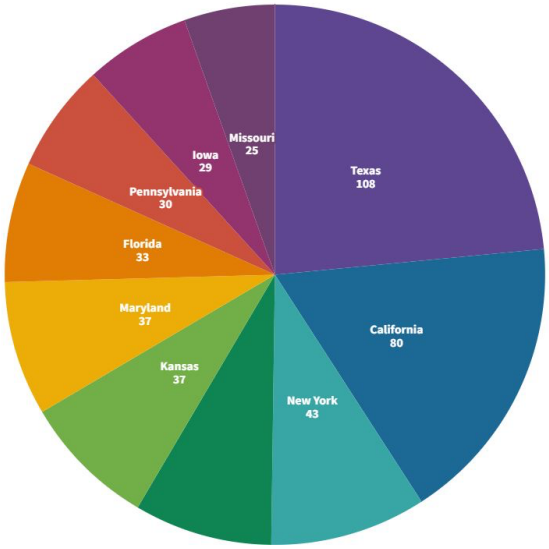
Data Completeness/Uniformity:
The following is the completeness and uniformity check that has been performed:
- The data is restricted to customers only from USA
- Every twitter data should have one and only one source and language
- A customer may have writer any number of tweets
- A customer may have any number of homes
- A customer can belong to maximum occupation

Ram Vaghani (NUID: 002704237)

Visualizations:



The above graph shows the average income of the default or not default customers for each maximum education degree they have received. We see that Higher education customers that are not default receive higher income and lower secondary education customers that are not default receive less income than they are default.



The top 10 states that customers own property at. Texas and California are two major states where customers own property.

Ram Vaghani (NUID: 002704237)

SQL to insert the data into database

- Used pymysql and mysql-connect libraries to connect and import the data files into the MySQL workbench

- Created the database using the below code:
  CREATE DATABASE dmdd;

- Created table using the below code:(an instance from our database)
  CREATE TABLE IF NOT EXISTS occupation (occupation_id int PRIMARY KEY NOT NULL, role varchar(45) NOT NULL)

- Inserted values to the table using the below code:(an instance from our database)
  INSERT INTO occupation (occupation_id,role) VALUES (%s,%s)

Ram Vaghani (NUID: 002704237)

Use Cases

1)Use Case: Count of customers for each occupation

Description: Giving the top occupations which are popular

Actor: User

Precondition: When a customer wants to buy something from shop, firstly he will be registered

Steps:

Actor action: User should have applied for credit

System Responses: System will provide the details of occupation if it is trustworthy or not.

Post Condition: Most popular occupation will receive credit

Alternate Path: The customer request is not correct and system throws an error

Error: User information is incorrect

```
1  select count(c.ï»¿sk_id) as count_of_customers, o.role
2  from creditloan c, occupation o
3  where c.occupation_id=o.ï»¿occupation_id
4  group by o.role
5  order by count(c.ï»¿sk_id) desc
6  limit 10;
7
8
9
10
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

count_of_customers	role
102	Laborers
87	Managers
47	Drivers
43	Core Staff
43	Sales Staff
23	High Skill Tech Staff
21	Accountants
16	Medicine staff
8	Private service staff
7	Security staff

Ram Vaghani (NUID: 002704237)

2) Use Case: People whose average income is more than rest of cities

Description: Average income of some cities are more compared to others and there is possibility to get credit faster compared to other cities

Actor: User

Precondition: the customer should be from one of the cities in which average income is high

Steps:

Actor action: User request for credit

System Responses: if the client fulfills all the prerequisite the loan will be santioned

Post Condition: Customer need to maintain high income

Alternate Path: there will be no alternate path

```
8  •  select avg(c.income) as avg_income, l.state
9     from creditloan c, location l
10    where c.ï»¿sk_id=l.sk_id
11    group by l.state
12    order by avg(income) desc
13    limit 10;
14
15
16
17
```

Result Grid | Filter Rows: | Export: | Wrap Cell Con

avg_income	state
301090.9091	Colorado
265263.1579	District of Columbia
259071.4286	Oklahoma
249750.0000	California
246093.7500	Virginia
240827.5862	Iowa
239931.8182	Florida
238860.0000	Missouri
238500.0000	Utah
234395.8333	Texas

Ram Vaghani (NUID: 002704237)

3)Use Case: which language the tweets were posted from the states

Description: Language popular in each state where number of tweets were more.

Actor: User

Precondition: Must have logged in the account

Steps:

System Responses: Tweet should be atleast one time

Post Condition: Highest number of tweets must be posted

Alternate Path: There will be no alternate path

```
15 •  select la.description,l.state, count(*) as count_of_tweets
16    from  twitter_data t, creditloan c, location l,language la
17    where t.sk_id=c.ï»¿sk_id
18    and c.ï»¿sk_id=l.sk_id
19    and la.ï»¿language_id=t.language_id
20    group by la.description, l.state;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

description	state	count_of_tweets
english	Illinois	4
french	Illinois	1
spanish	Wisconsin	1
spanish	Texas	3
chinese	North Carolina	1
english	North Carolina	1
spanish	North Carolina	1
english	California	17
english	Texas	33
english	Indiana	1
spanish	Indiana	1
chinese	California	1
english	Nebraska	5
english	Washington	7
french	Texas	1

Ram Vaghani (NUID: 002704237)

4)Use Case: What is the source in which tweets were posted in each state.

Description: Source popular in each state where number of tweets were more.

Actor: Title

Precondition: Must have logged in the account

Steps:
System Responses: Tweet should be posted more than one time

Post Condition: Highest number of tweets must be posted

Alternate Path: There will be no alternate path

```
22 •  select s.title,l.state, count(*) as count_of_tweets
23    from  twitter_data t, creditloan c, location l,source s
24    where t.sk_id=c.ï»¿sk_id
25    and c.ï»¿sk_id=l.sk_id
26    and s.ï»¿source_id=t.source_id
27    group by s.title, l.state;
28
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

title	state	count_of_tweets
Hootsuite Inc.	Illinois	1
Twitter Web Client	Illinois	4
Twitter Web Client	Wisconsin	3
Twitter Web Client	Texas	27
Twitter Web Client	North Carolina	3
Hootsuite Inc.	California	7
Twitter Web Client	California	12
Hootsuite Inc.	Texas	13
Hootsuite Inc.	Indiana	1
Twitter Web Client	Indiana	1
Hootsuite Inc.	Nebraska	2
Twitter Web Client	Nebraska	3
Hootsuite Inc.	Washington	2
Twitter Web Client	Washington	6
Hootsuite Inc.	Virginia	3

Ram Vaghani (NUID: 002704237)

5) Use Case: The average income for occupation in each case of default or not default
Description: User views the orders made by him/her

Actor: User

Precondition: Customer has either been defaulted or not defaulted

Steps:
System Responses: Displays all the income according to the role

```
29 •  select o.role, c.target, avg(c.income) as average_income
30    from occupation o, creditloan c
31    where c.occupation_id=o.ï»¿occupation_id
32    group by o.role, c.target;
33
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

role	target	average_income
Laborers	0	200571.4286
Managers	0	305470.5882
Laborers	1	196875.0000
Drivers	0	247600.0000
Core Staff	0	214123.1707
Sales Staff	0	227013.1579
Sales Staff	1	181800.0000
High Skill Tech Staff	0	205977.2727
Medicine staff	0	191892.8571
Accountants	0	258000.0000
Private service staff	1	90000.0000
Medicine staff	1	135000.0000
Private service staff	0	192857.1429
Cooking staff	0	166500.0000
HR staff	0	135000.0000
High Skill Tech Staff	1	265500.0000
Cleaning staff	1	157500.0000