Data cleaning

TABLE 1 “accountsdata” -> data cleanup done

1. **Overview of the "accountsdata" Table:**

The "accountsdata" table comprises 5 columns and a total of 1054 rows of data.

The columns are defined as follows:

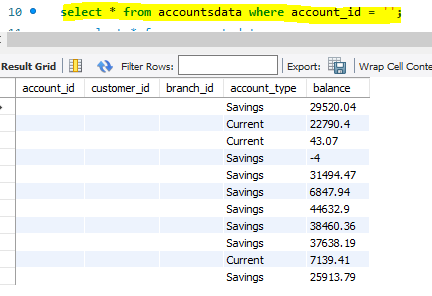
1. account\_id -> int
2. customer id -> int
3. branch\_id -> int
4. account\_type -> varchar(20)
5. balance -> double

**2. Discrepancy Identified in the Data:**

During the data review process, discrepancies were detected in the dataset. These discrepancies were primarily focused on the presence of missing or invalid values.

3**. Rows with Missing Values:**

Upon further investigation, 11 rows were found to be missing account\_id, customer\_id, and branch\_id.



**Impact of the Issue:**

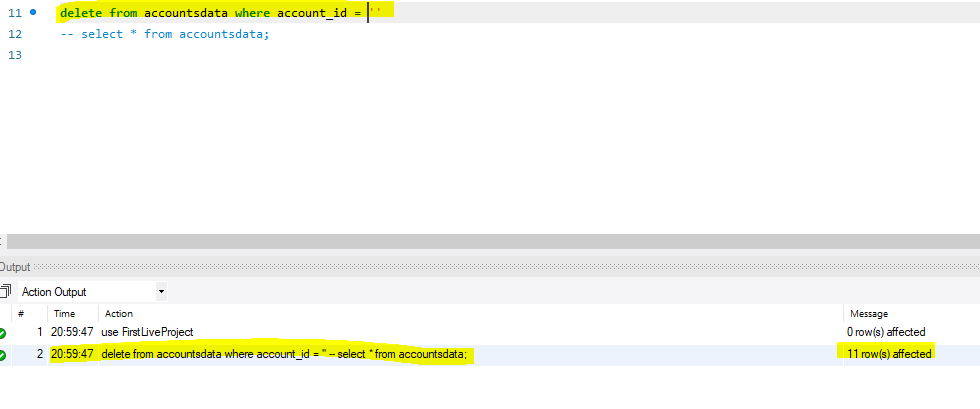
Both account\_id and customer\_id are mandatory fields and cannot be null, as they are integral to identifying unique records in the table.

Since the integrity of the table relies on these fields being populated, any record with missing values in these columns cannot be considered valid.

**Action Taken:**

As the account\_id and customer\_id must always be unique and non-null, the 11 rows with missing values for these fields were identified for removal.

These rows were deleted to ensure that only valid and complete records remained in the dataset.



4. **Removal of Blank Values:**

After the deletion of the 11 rows with missing account\_id, customer\_id, and branch\_id, the table was free of any blank values.

5. **Identification of Duplicate Rows:**

Additionally, 43 duplicate rows were identified during the data cleaning process. These duplicates were determined to be exact repetitions of other records in the table.



Removed duplicate values by coping distinct data from old table.



**Action Taken:**

These duplicate rows were removed to maintain data integrity and avoid redundancy.

**6. Final Cleaned Data:**

After addressing the missing values and duplicates, the table now contains a total of 1000 valid rows of data. This data is now clean, consistent, and ready for further analysis or use.

**7. Conclusion:**

The cleaning process involved:

Removing rows with missing account\_id, customer\_id, and branch\_id (**11 rows**).

Deleting duplicate records (43 rows).

Resulting in a clean dataset of 1000 rows.

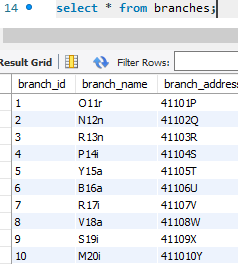
TABLE 2 “branches” -> data cleanup done

1. **Overview of the "branches" Table:**

* The "branches" table consists of 3 columns and a total of 10 rows of data.
* The columns are structured as follows:
  1. branch\_id – INT
  2. branch\_name – varchar(20)
  3. branch\_address – varchar(40)

2. **Discrepancy Review:**

* A thorough review and analysis of the data in the "branches" table was conducted.
* During this review, no discrepancies were found within the table. All data fields were complete, accurate, and consistent.



**3. Conclusion:**

* Since no discrepancies or issues were detected, there was no need for any data cleaning.
* The table is already in a clean, valid state and is ready for further analysis or use without requiring any modifications.

TABLE 3 “customersdata” -> data cleanup done

1. Overview of the "customersdata" Table:

* The "customersdata" table includes 6 columns and 1134 rows in total.
* The columns in the table are defined as follows:
  1. customer\_id – int
  2. first\_name – varchar(20)
  3. last\_name – varchar(20)
  4. email – varchar(50)
  5. phone – varchar(50)
  6. address – varchar(200)

2. Discrepancy Review:

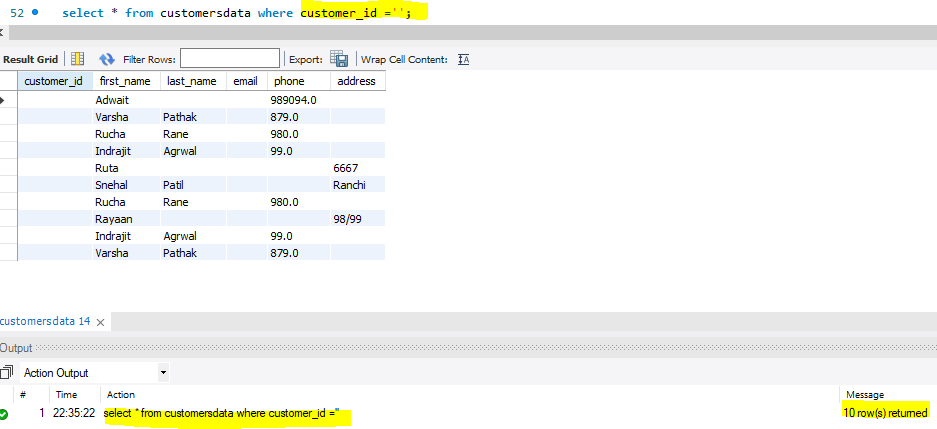
A detailed analysis of the data in the "customersdata" table was carried out, revealing several discrepancies within the raw data that required attention.

* The following issues were identified:
  1. 10 rows with missing customer\_id.
  2. 2 rows with missing first\_name.
  3. 7 rows with missing last\_name.
  4. 15 rows with missing email.
  5. 5 rows with missing phone.
  6. 1 row with missing address.

3. Data Cleaning Actions: To maintain the integrity of the dataset, it was necessary to address these issues. Specifically, rows with missing critical data (i.e., customer\_id, first\_name, last\_name, email, phone, or address) were flagged for removal, as they are essential for valid customer records. As a result, the rows with incomplete information were removed from the dataset.

Additionally, duplicate and blank rows were eliminated, ensuring that only unique, valid records remained.

Found 10 rows with no customer\_id

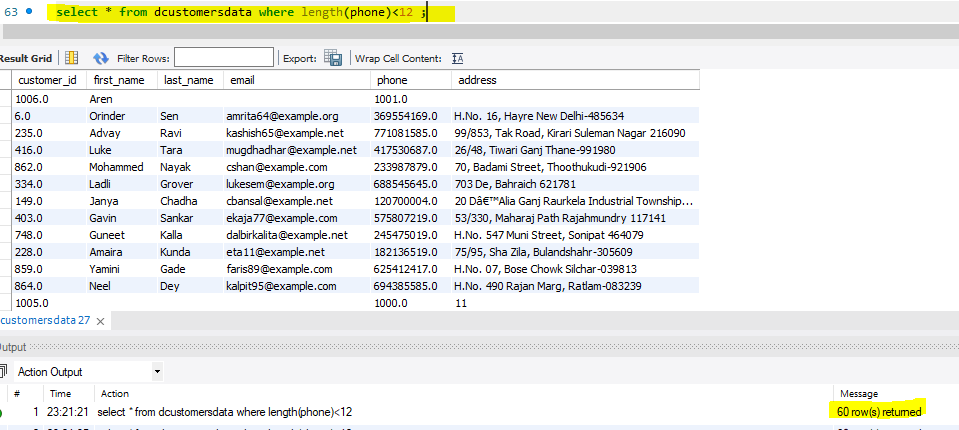


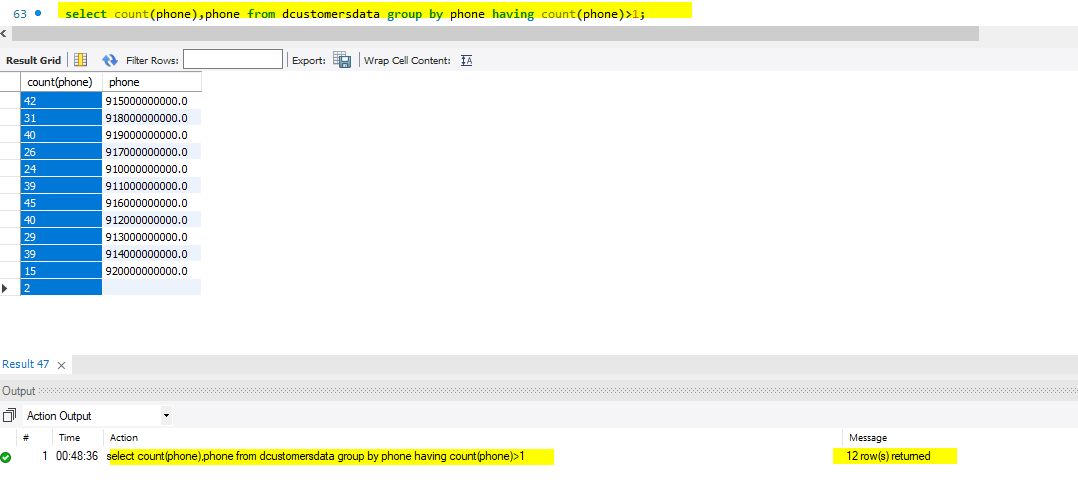


Removed duplicate and kept distinct data since entire row was having same data.

Removed duplicate / blank rows from table



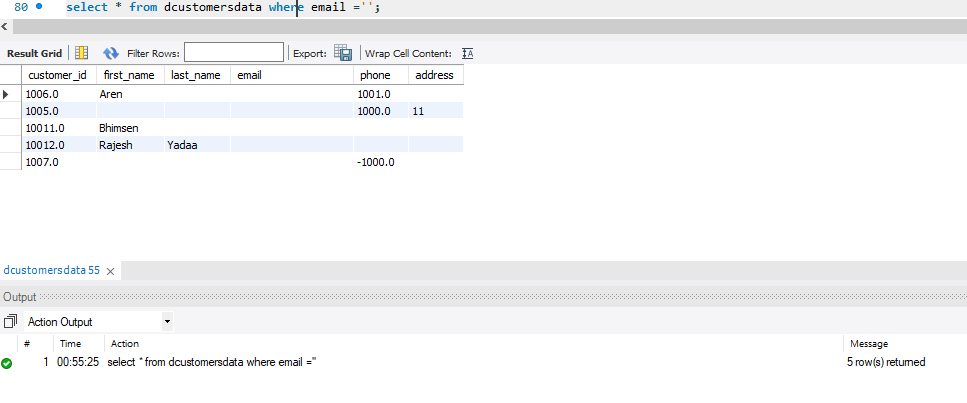


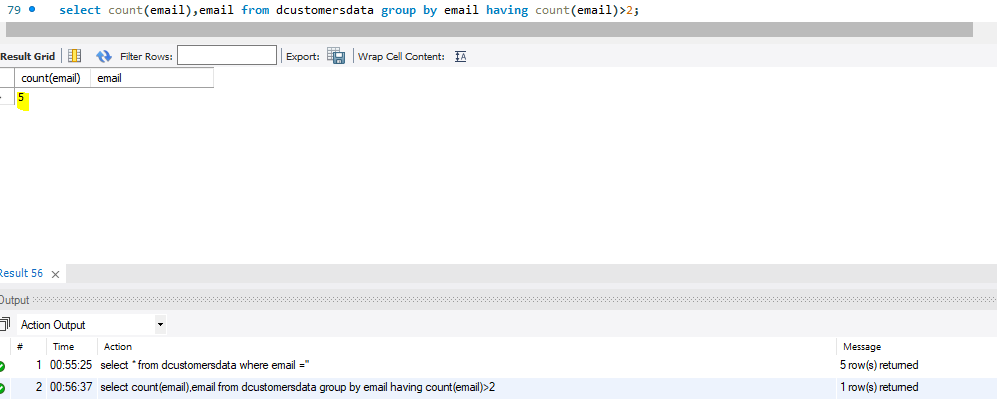


As per discussion with Sarang sir, we entered the dummy phone numbers.

Total of duplicate phone 372

Found 5 rows which does not have email id’s.





**Conclusion**: Following these cleaning procedures, the dataset was thoroughly refined. The "customersdata" table now contains 1000 valid records, free of discrepancies, and is fully prepared for further analysis or processing.

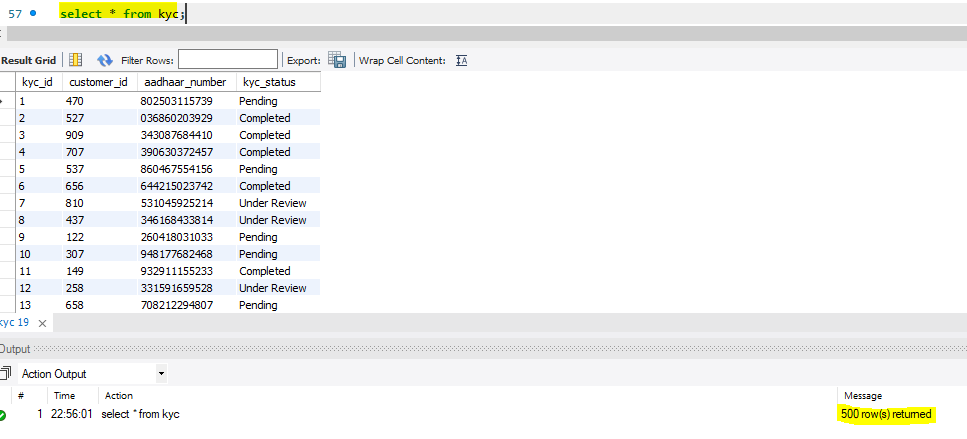
TABLE 4 “kyc” -> data cleanup done

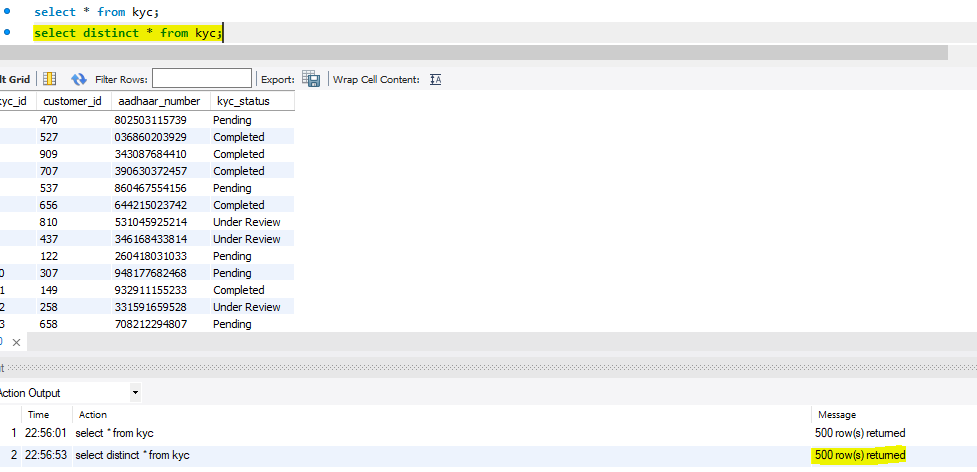
**1. Overview of the "kyc" Table:**

* The "kyc" table includes 4 columns and a total of 500 rows of data.
* The columns in the table are defined as follows:
  1. kyc\_id – INT
  2. customer\_id – INT
  3. aadhaar\_number – TEXT
  4. kyc\_status – TEXT

**2. Discrepancy Review:**

* A thorough review of the "kyc" table was conducted to identify any potential discrepancies or issues with the data.
* No discrepancies were found in the table. All data fields were complete, accurate, and consistent.





**3. Conclusion:**

* Since there were no discrepancies detected in the "kyc" table, no data cleaning was required.
* The table is already in a clean and valid state, making it ready for further analysis or use.

TABLE 5 “loans1” -> data cleanup done

14 rows does not have loan\_id however it is having customer id so we are not deleting data. Seems it is genuine.

* 1. Table includes 6 columns ( loan\_id, customer\_id, loan\_type, loan\_amount, interest\_rate, loan\_status ) with 564.
     1. loan\_id -> varchar(10)
     2. customer\_id -> int
     3. loan\_type -> varchar(10)
     4. loan\_amount -> double
     5. interest\_rate -> varchar(10)
     6. loan\_status -> varchar(10)
  2. Discrepancy found in table.
  3. Found 14 rows found with no loan\_id in raw data.
  4. Found 15 rows found with no customer\_id in raw data.
  5. Found 15 rows found with no interest\_rate in raw data.

Found 1 row found which is only having loan\_amount in raw data

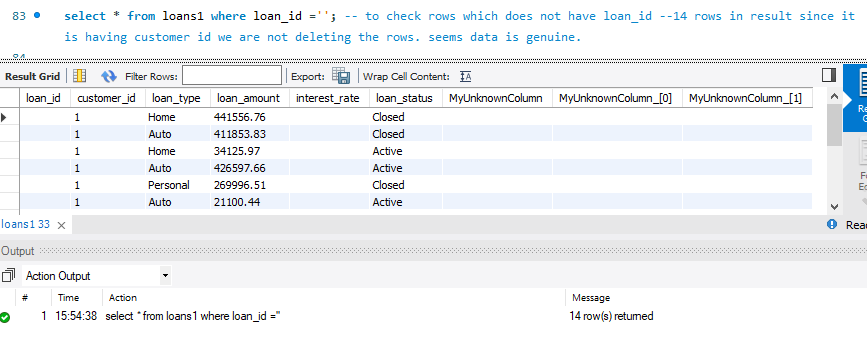
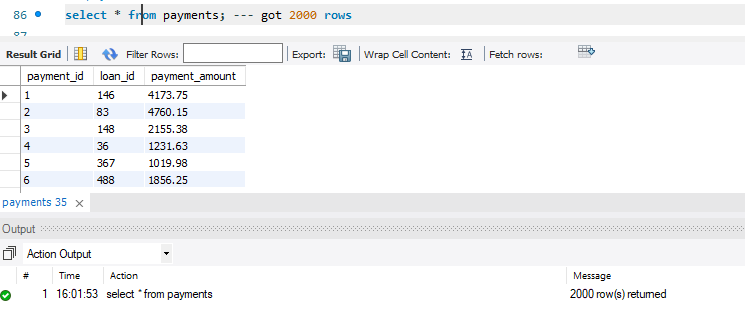


TABLE 6 “payments” -> data cleanup done

1. Table includes 3 columns ( payment\_id, loan\_id, payment\_amount ) with 2000 rows.
   * 1. payment\_id -> int
     2. loan\_id -> int
     3. payment\_amount -> double
2. No discrepancy found in table



Found all 2000 rows are distinct hence data is already clean.

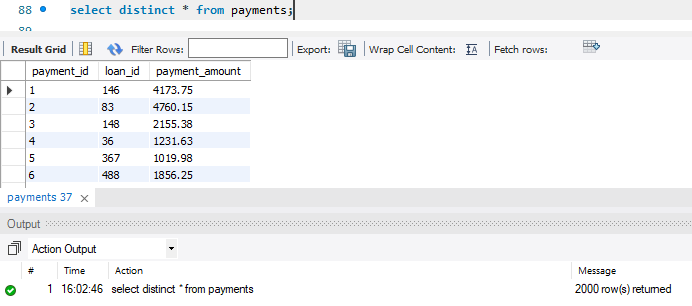
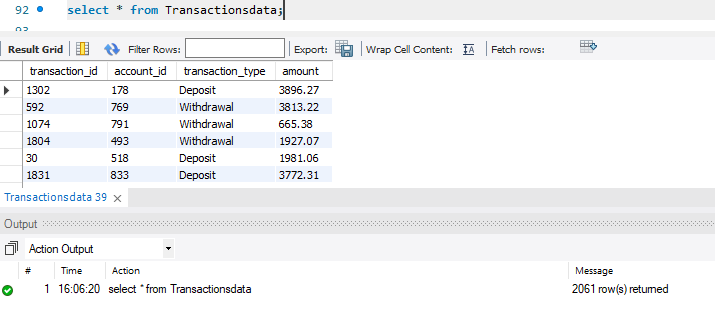


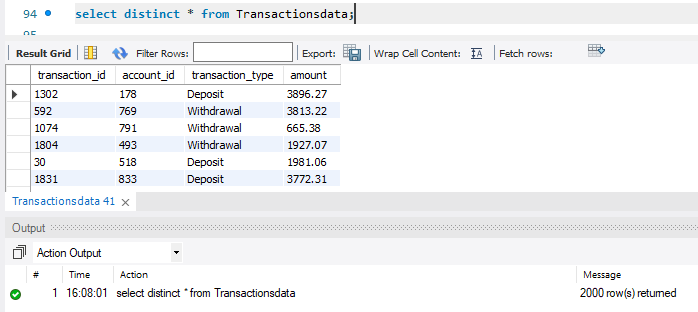
TABLE 7 “Transactionsdata” -> data cleanup done

Table includes 4 columns ( transaction\_id, account\_id, transaction\_type, amount ) with 2061 rows and all are imported in sql.

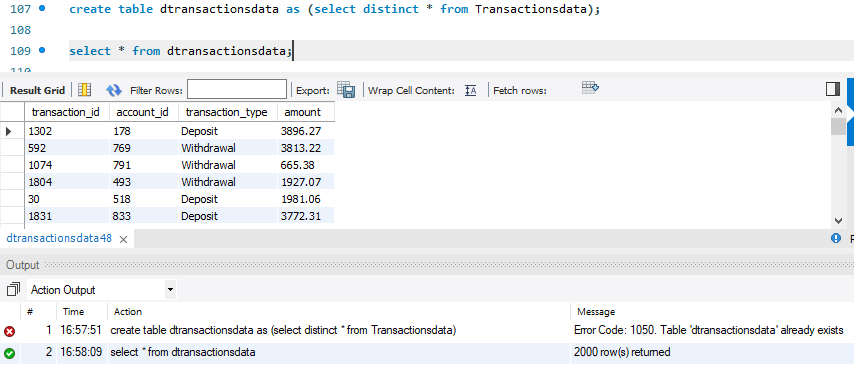
* + 1. transaction\_id -> int
    2. account\_id -> int
    3. transaction\_type -> varchar(20)
    4. amount -> double

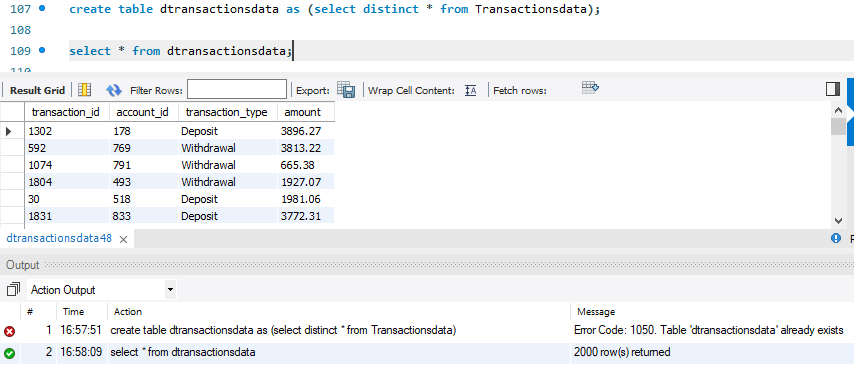
1. No discrepancy found in table





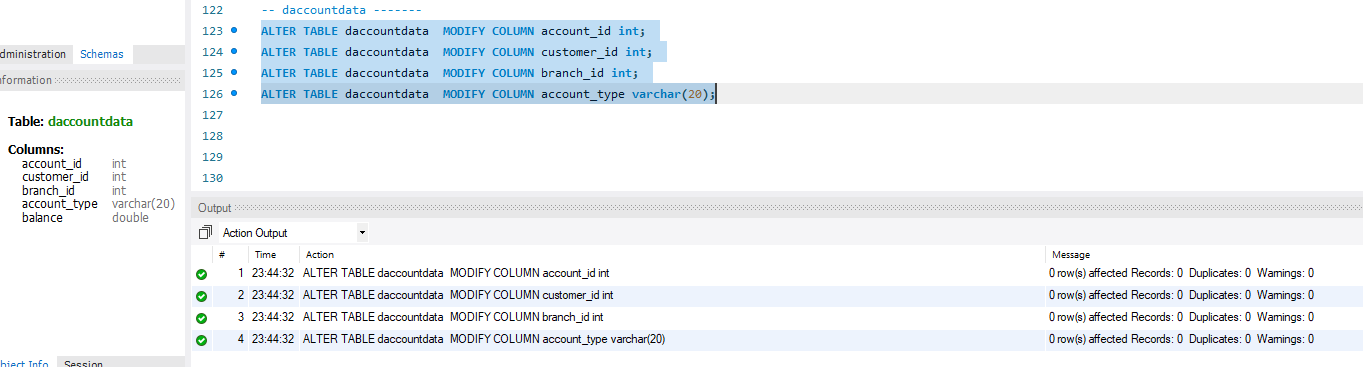
Removed duplicate and kept distinct data since entire row was having same data.



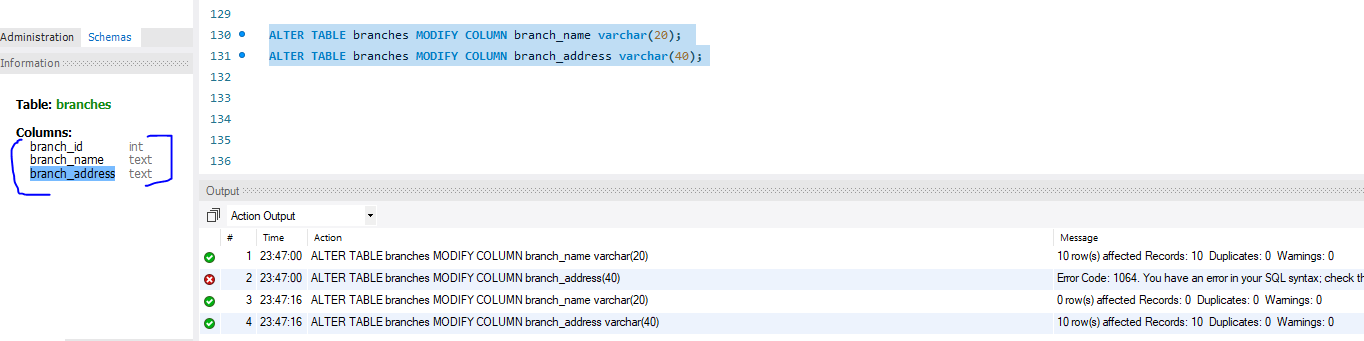


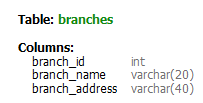
Data type change -🡪 need to work on document part from here

Daccountdata

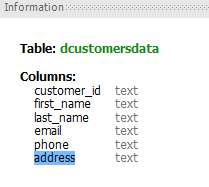


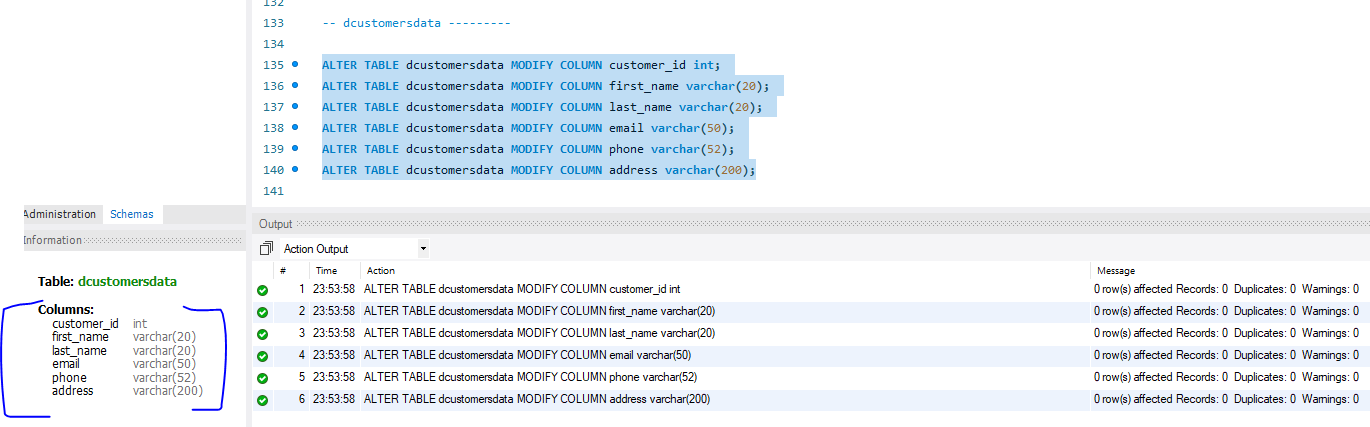
Branches





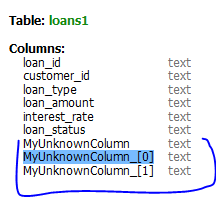
Dcustomersdata

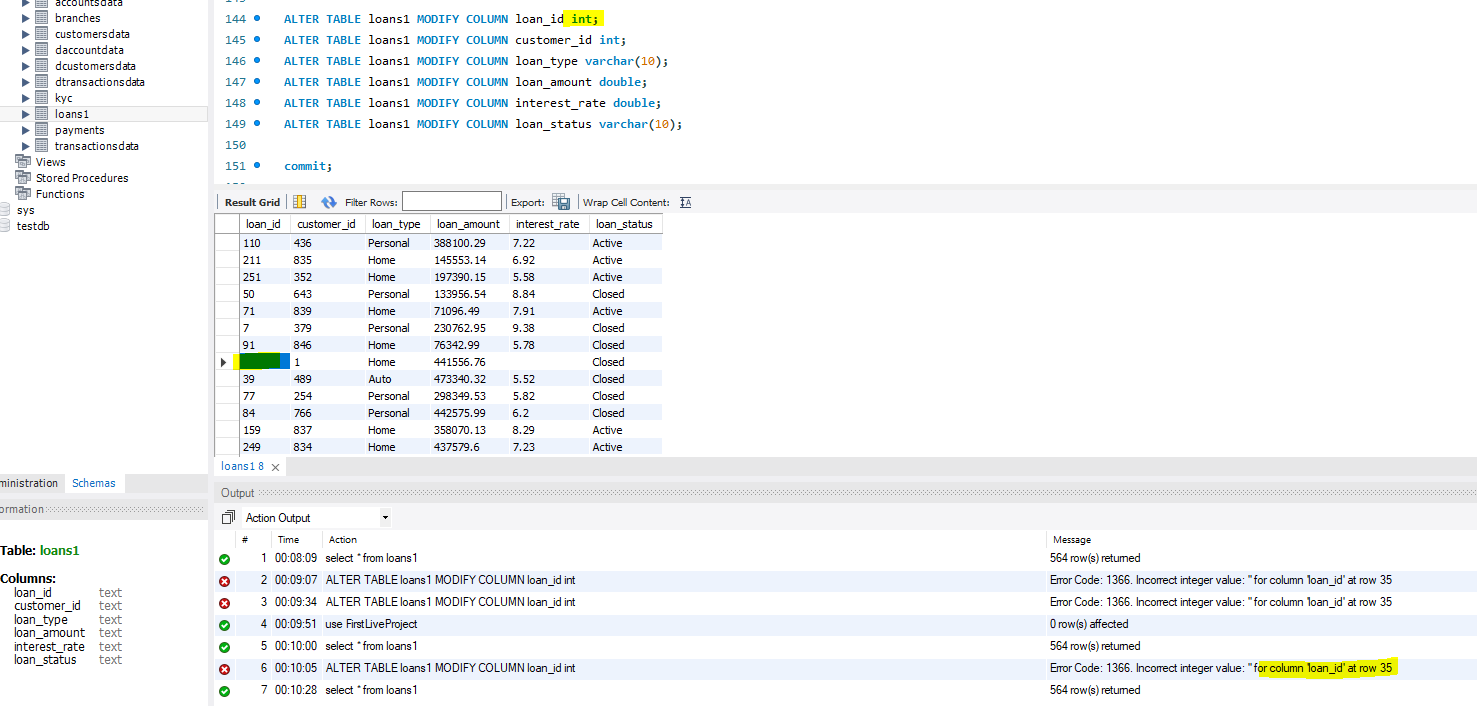




Loans1

Unknown column found







Payments

