FINAL PROJECT REPORT

ISM 6218 – Advanced Database Management

Submitted by Group 4

Uma Srikanth Reddy Koduru	U94125452
Prasanna Eddala	U24108900
Mohan Vamsi Karasala	U52431177
Akhilasri Musikara	U82181510

Major in

BUSINESS ANALYTICS AND INFORMATION SYSTEMS

Under the guidance of

DR. DONALD BERNDT

MUMA COLLEGE OF BUSINESS

UNIVERSITY OF SOUTH FLORIDA

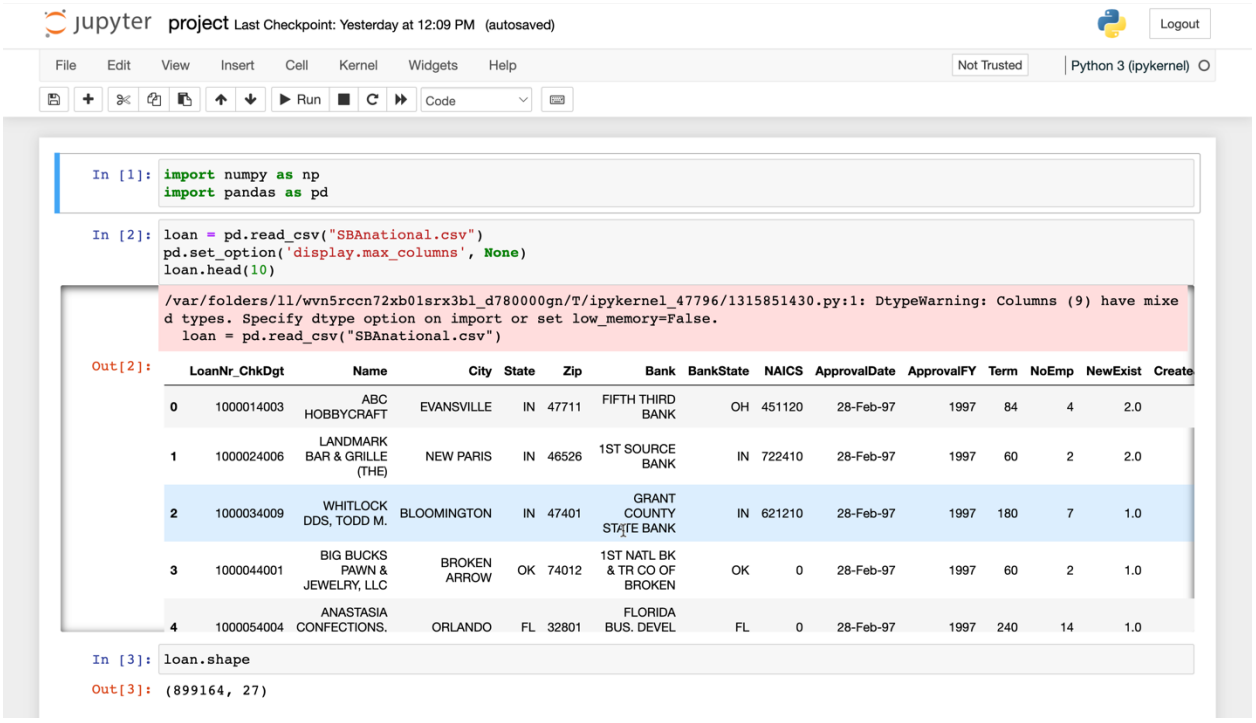
TOPIC AREA	DESCRIPTION	POINTS	TEAM MEMBERS
Database Design	This part should include a logical database design (for the relational model), using normalization to control redundancy and integrity constraints for data quality.	**25**	**Srikanth** **Prasanna** **Mohan** **Akhila**
Query Writing	This part is another chance to write SQL queries, explore transactions, and even do some database programming for stored procedures.	**25**	**Srikanth** **Prasanna** **Mohan** **Akhila**
Performance Tuning	In this section, you can capitalize and extend your prior experiments with indexing, optimizer modes, partitioning, parallel execution and any other techniques you want to further explore	**25**	**Srikanth** **Prasanna** **Mohan** **Akhila**
Other Topics	Here you are free to explore any other topics of interest. Suggestions include DBA scripts, database security, interface design, data visualization, data mining, and NoSQL databases.	**25**	**Srikanth** **Prasanna** **Mohan** **Akhila**

Table of Contents

Introduction

The data set is about US Small Business Administration and about the banks that have lend to small businesses over a period. A database has been created using this database and visualizations are done to present it to the end user. The technologies used in the project are Python, Oracle SQL Developer, Excel, and Tableau.

First the data is loaded and cleaned using python.

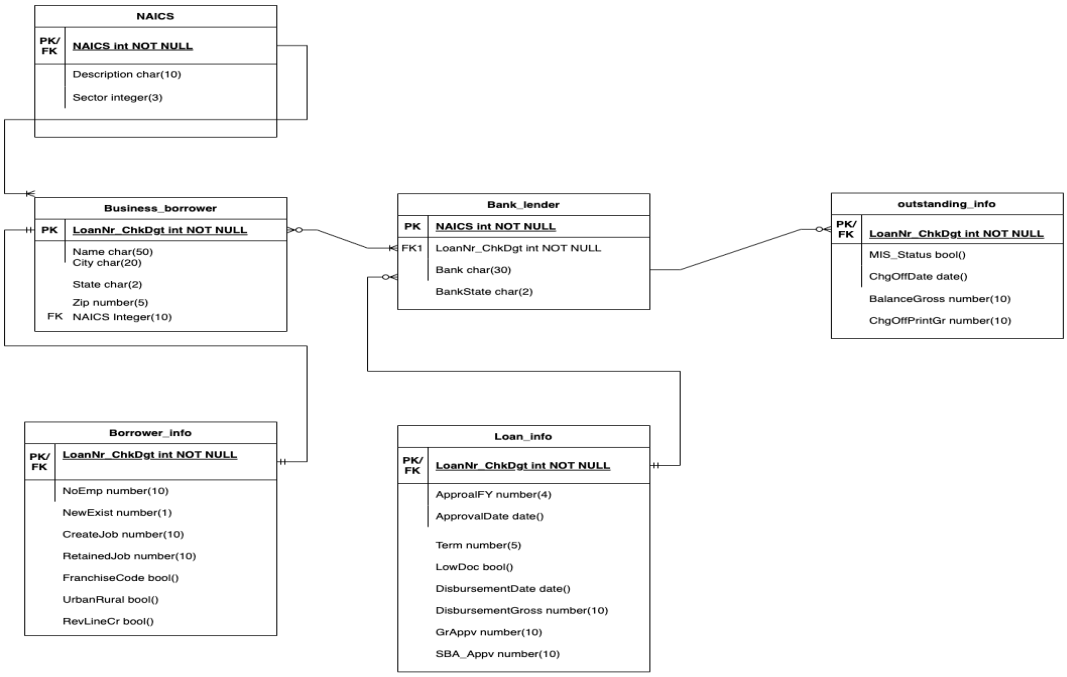


CODE:

```
import numpy as np
import pandas as pd
loan = pd.read_csv("SBAnational.csv")
pd.set_option('display.max_columns', None)
loan.head(10)
loan.shape
loan.drop_duplicates(inplace = True)
loan.shape
loan = loan[loan['NAICS']!=0]
loan.shape
unique_cols = [col for col in loan.columns if not loan[col].duplicated().any()]
unique_cols
loan['NAICS'].value_counts()
loan.to_csv('loan_data.csv', index=False)
```

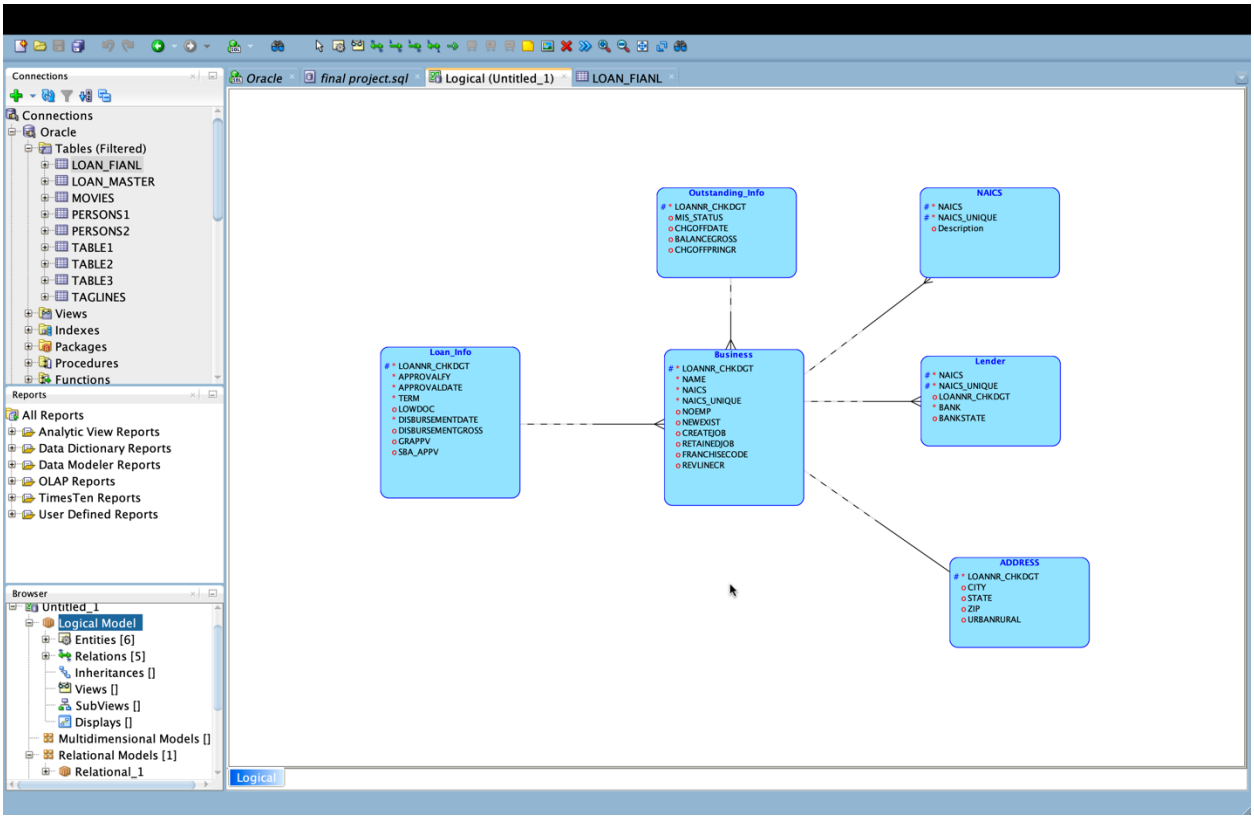
Part 1: Database Design

WITHOUT NORMALIZATION:

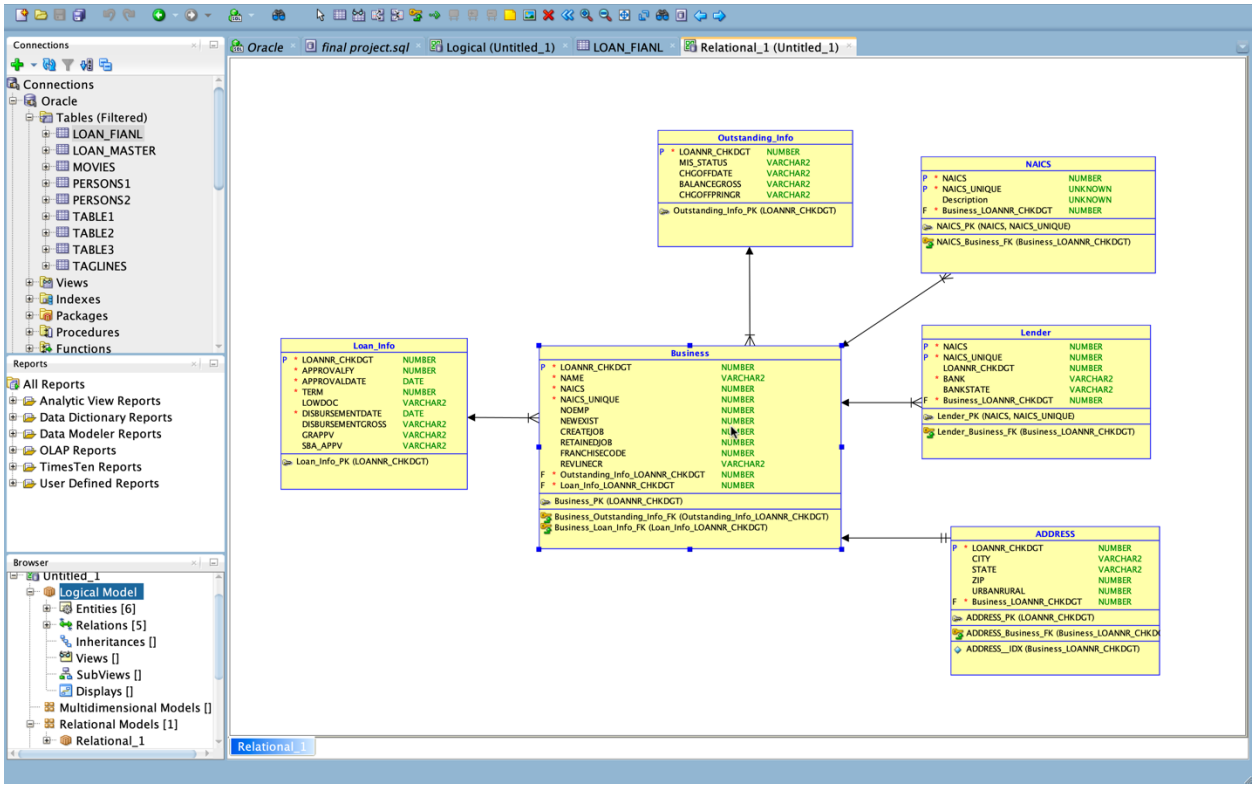


AFTER NORMALIZATION:

The rules of 1NF are followed from the python code, the business_borrower and Borrower_Info tables have been merged to make the design 3NF compliant (No transitive dependencies). The design for the normalized tables is given below. In the below logical diagram all the relations between the tables have been defined and primary keys are set.



The logical diagram is further converted into a relational diagram.



Section 1.1 Data Integrity:

ALTER TABLE LOAN_FIANL
ADD CONSTRAINT naics_as_unique UNIQUE (NAICS_UNIQUE);

ALTER TABLE BUSINESS
ADD CONSTRAINT loannr_chkdgt PRIMARY KEY (LOANNR_CHKDGT);

ALTER TABLE lENDER
ADD CONSTRAINT LENDER_PK PRIMARY KEY (NAICS,NAICS_UNIQUE);

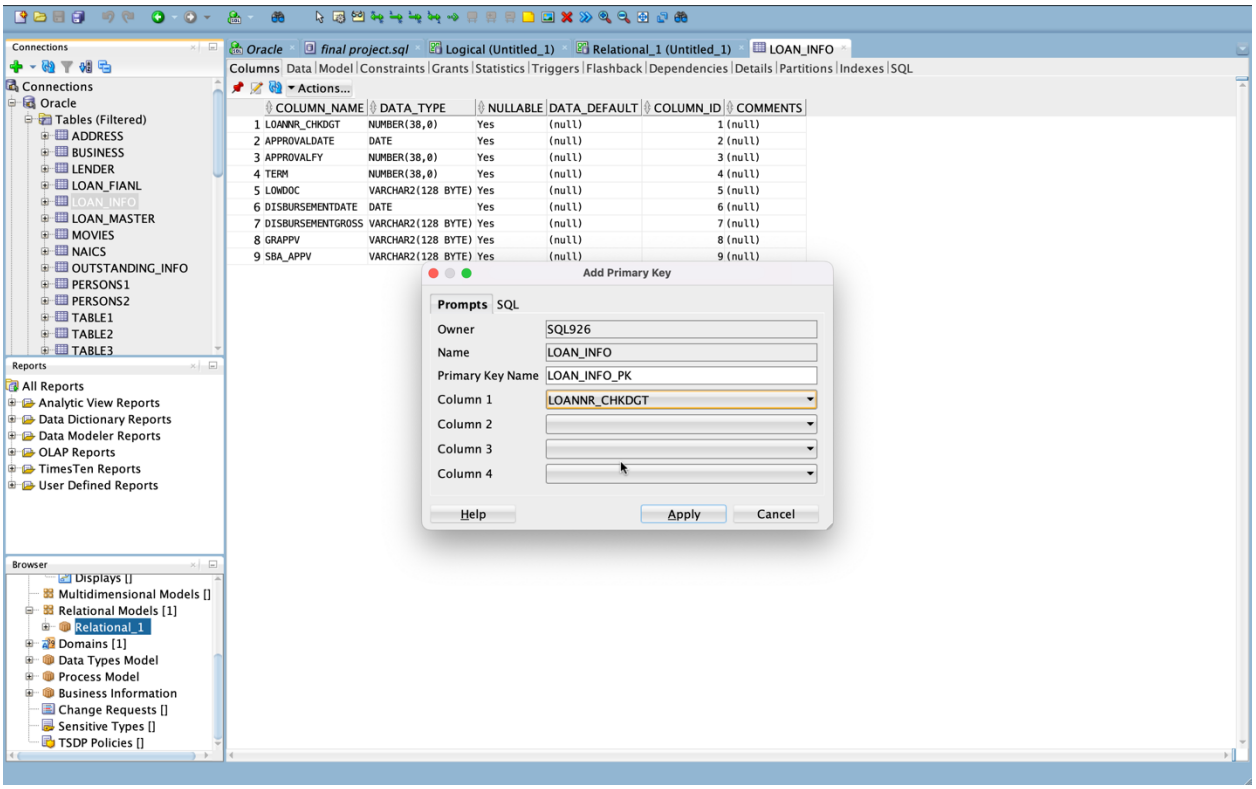
ALTER TABLE NAICS
ADD CONSTRAINT NAICS_PK PRIMARY KEY (NAICS,NAICS_UNIQUE);

ALTER TABLE OUTSTANDING_INFO
ADD CONSTRAINT OUTSTANDING_PK PRIMARY KEY (LOANNR_CHKDGT);

ALTER TABLE BUSINESS

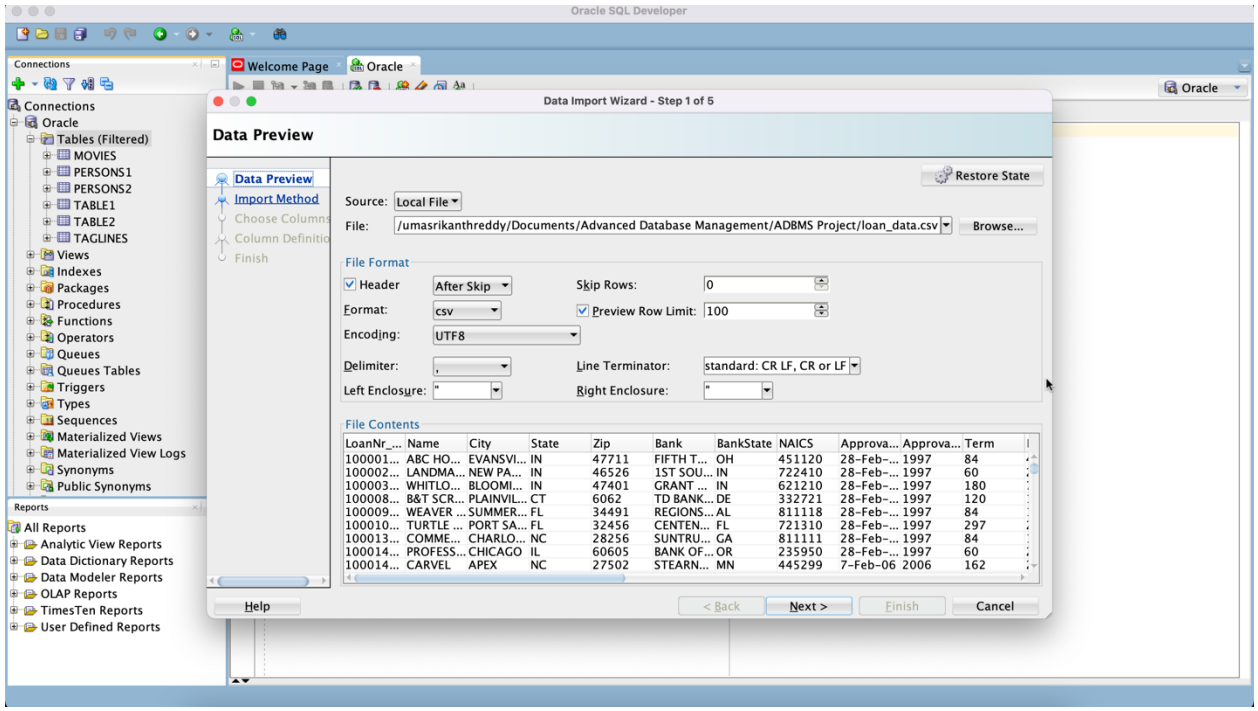
ADD CONSTRAINT LOAN_INFO_PK PRIMARY KEY (LOANNR_CHKDGT);

All the primary key constraints have been added to the tables.

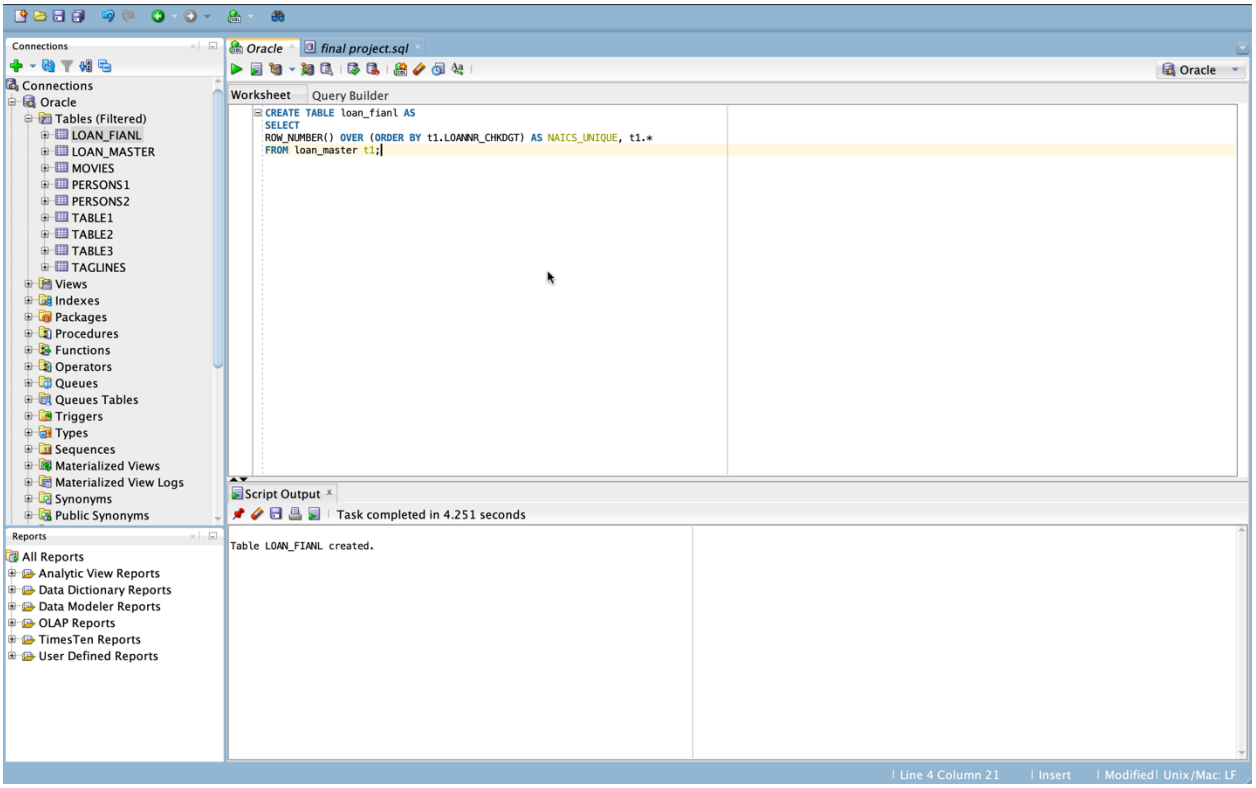


Section 1.2: Data Generation and Loading:

Tables are created using the csv file obtained from the python code.



The NAICS table did not have a unique column in the raw data, so a new column (NAICS_UNIQUE) is created using the code given below. This column together with NAICS column would serve as primary key in the design.


Connections
Oracle
final project.sql
Worksheet
Query Builder
CREATE TABLE loan_fianl AS
SELECT
ROW_NUMBER() OVER (ORDER BY t1.LOANNR_CHKDGT) AS NAICS_UNIQUE, t1.*
FROM loan_master t1;
Tables (Filtered)
LOAN_FIANL
LOAN_MASTER
MOVIES
PERSONS1
PERSONS2
TABLE1
TABLE2
TABLE3
TAGLINES
Views
Indexes
Packages
Procedures
Functions
Operators
Queues
Queues Tables
Triggers
Types
Sequences
Materialized Views
Materialized View Logs
Synonyms
Public Synonyms
Reports
All Reports
Analytic View Reports
Data Dictionary Reports
Data Modeler Reports
OLAP Reports
TimesTen Reports
User Defined Reports
Script Output
Task completed in 4.251 seconds
Table LOAN_FIANL created.
Line 4 Column 21
Insert
Modified
Unix/Mac: LF

Part 2: Query Writing

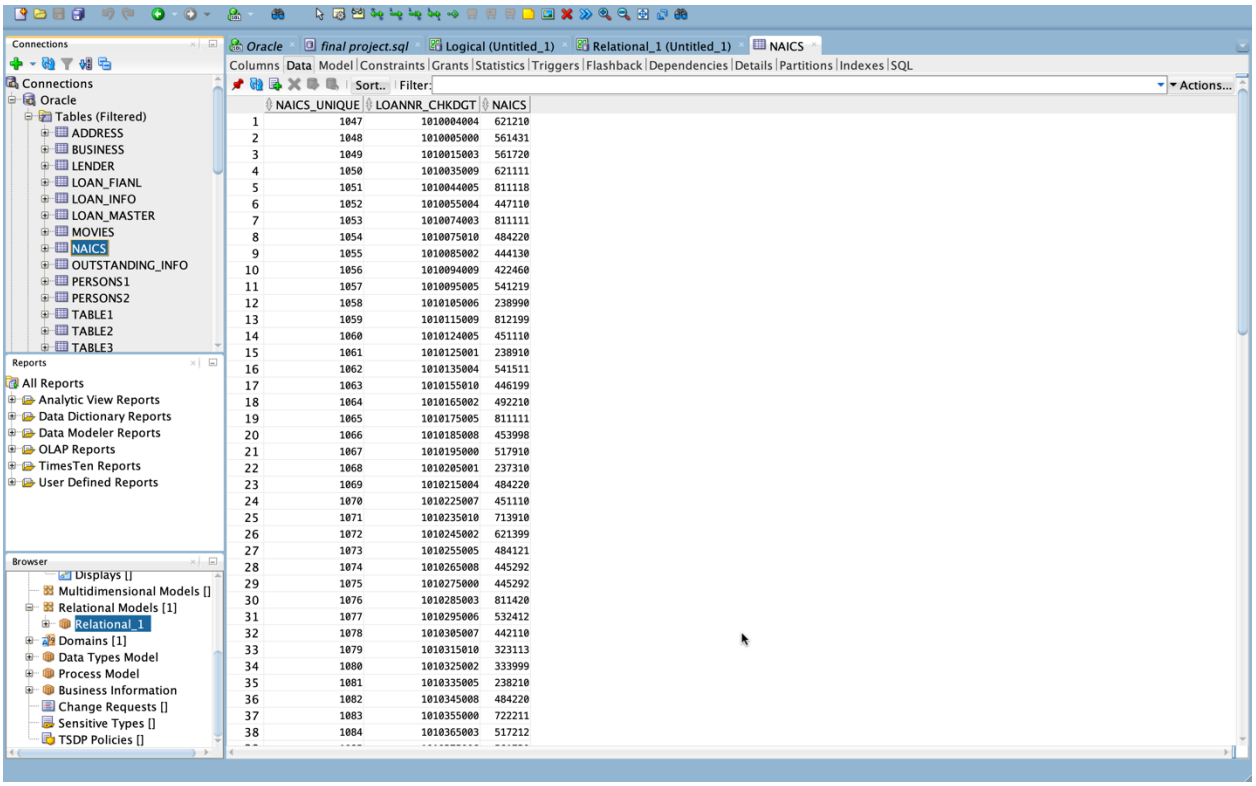
Section 2.1: Database Programming:

A stored procedure (create_table) has been created to create a new table with NAICS_UNIQUE column to fill it with unique numbers.

```
CREATE OR REPLACE PROCEDURE create_table AS
BEGIN
  EXECUTE IMMEDIATE 'CREATE TABLE loan_final AS SELECT ROW_NUMBER() OVER (ORDER BY
t1.LOANNR_CHKDGT) AS NAICS_UNIQUE, t1.* FROM loan_master t1';
END;
/

EXECUTE create_table;
```

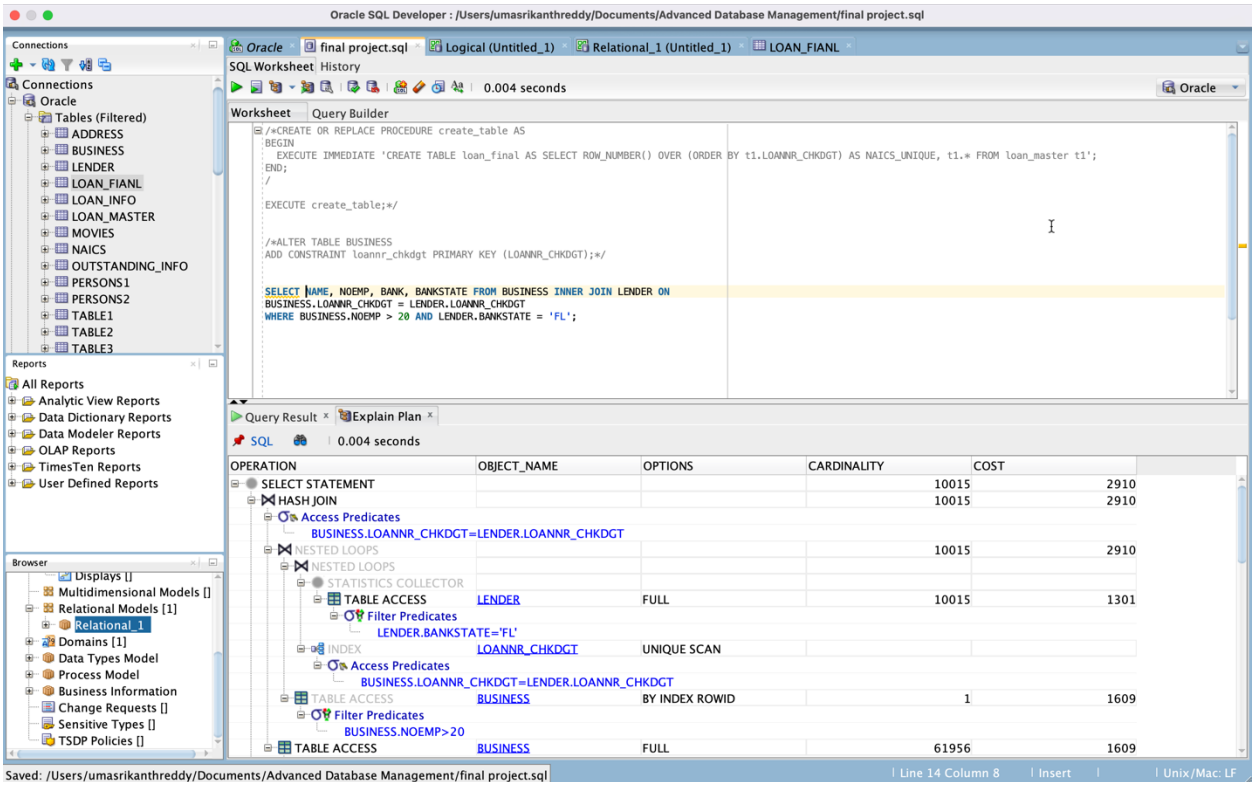
The unique rows are shown in the below picture.



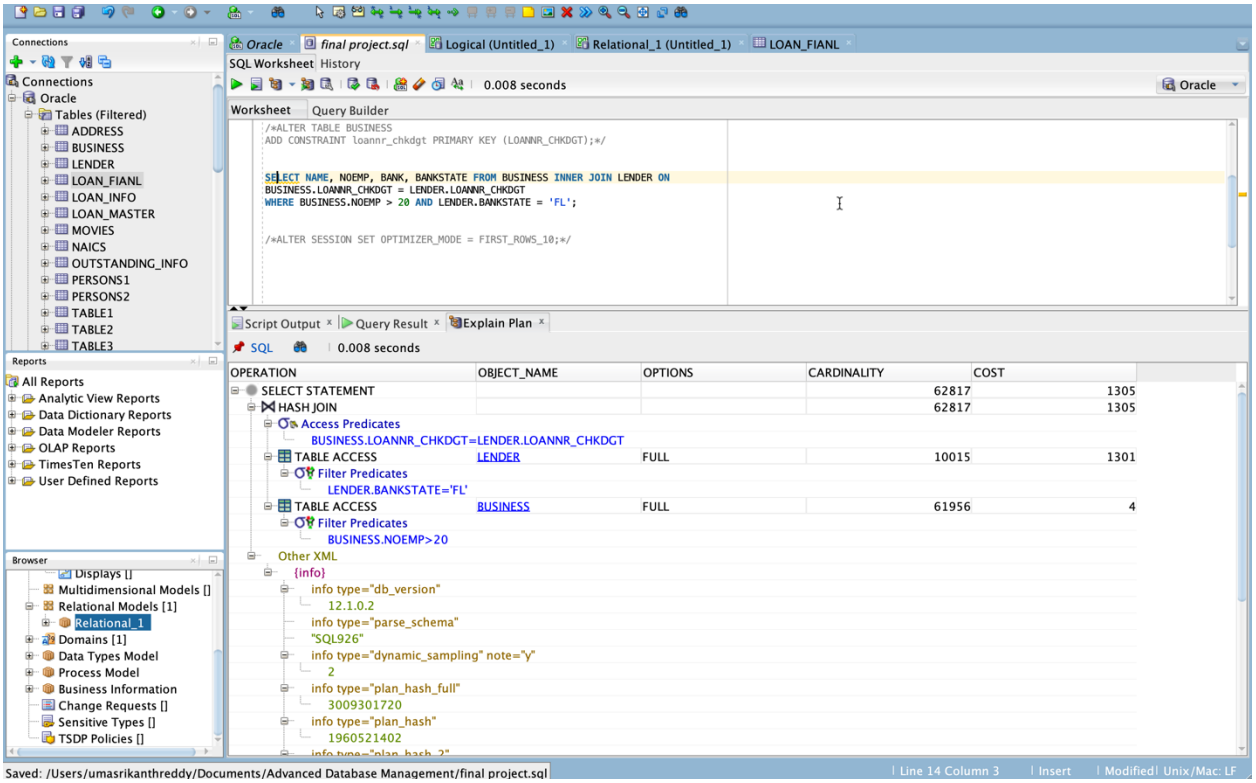
Part 3: Performance Tuning

TUNING 1:

(1) the purpose of the experiment is to change the optimizer mode so as to get in the query results with least computational power.
(2) A select statement is executed and the cost is noted using the default optimizer mode, then the optimizer mode is set to first 100 rows and the cost is noted.

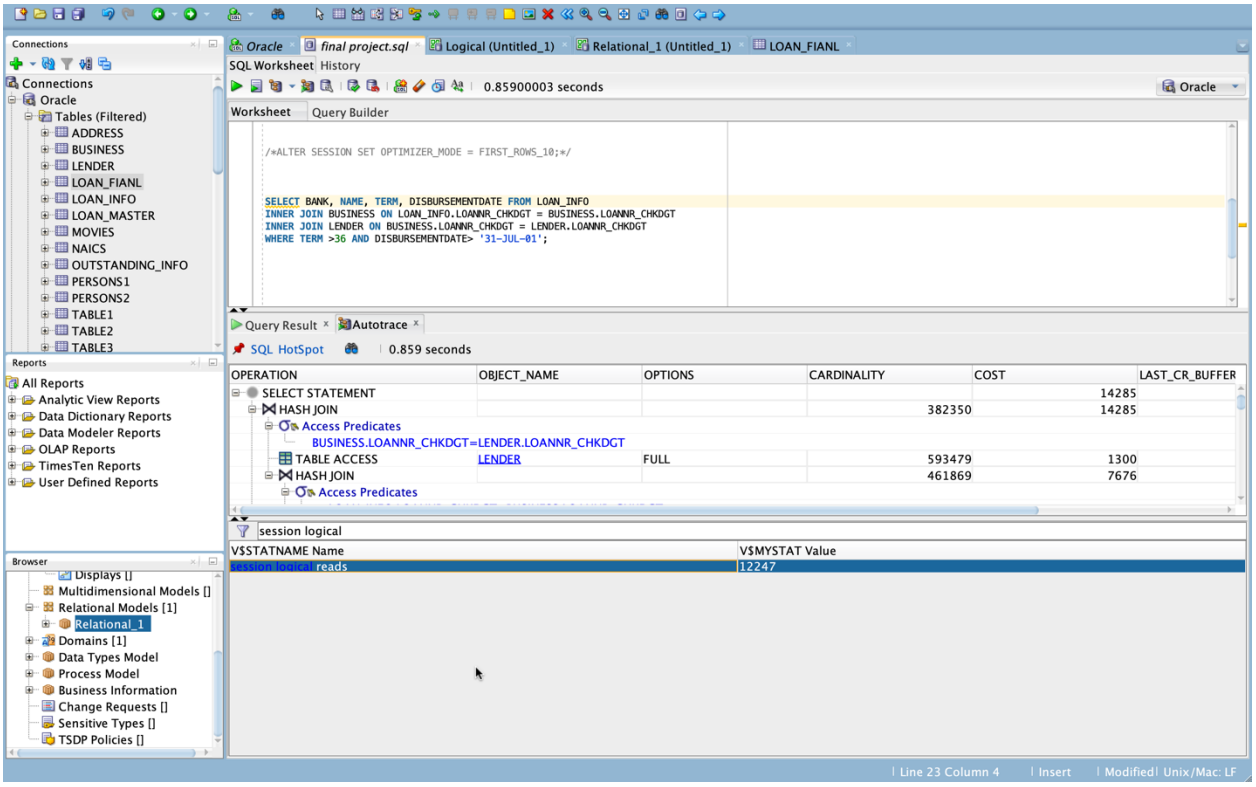


(3) There is an improvement in the cost as shown below

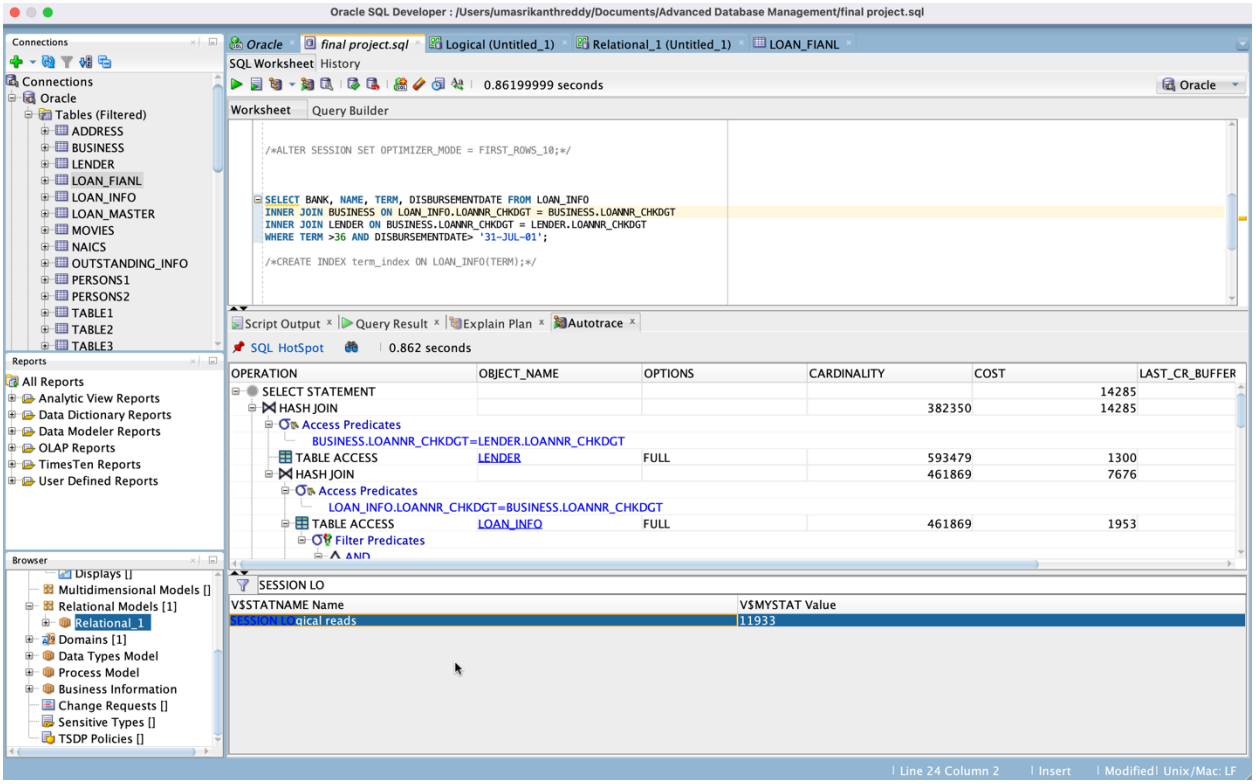


TUNING 2:

(1) The purpose of the below tuning operation is to check wheter index has worked or not to reduce the cost.
(2) A select statement is executed and the cost, Session logical reads are noted, then an index (term_index) is created and the cost are noted.



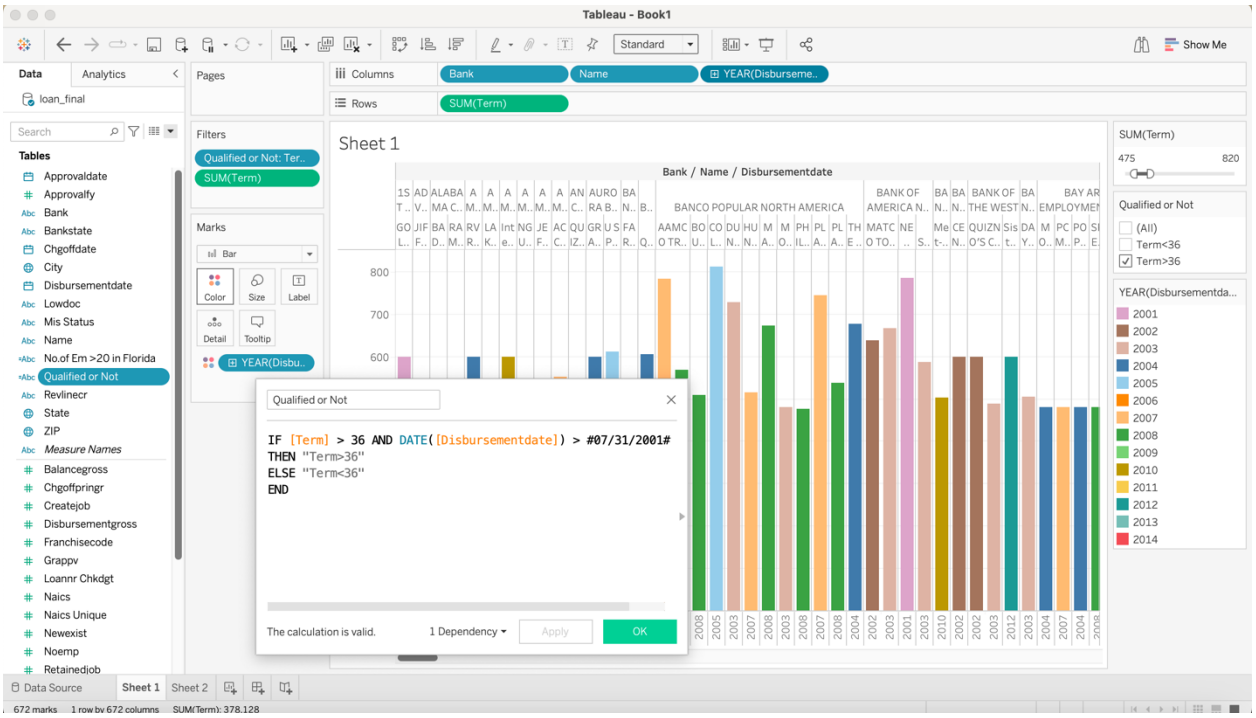
(3) There is no change in the cost, but the session logical reads are reduced after creating the index as shown below



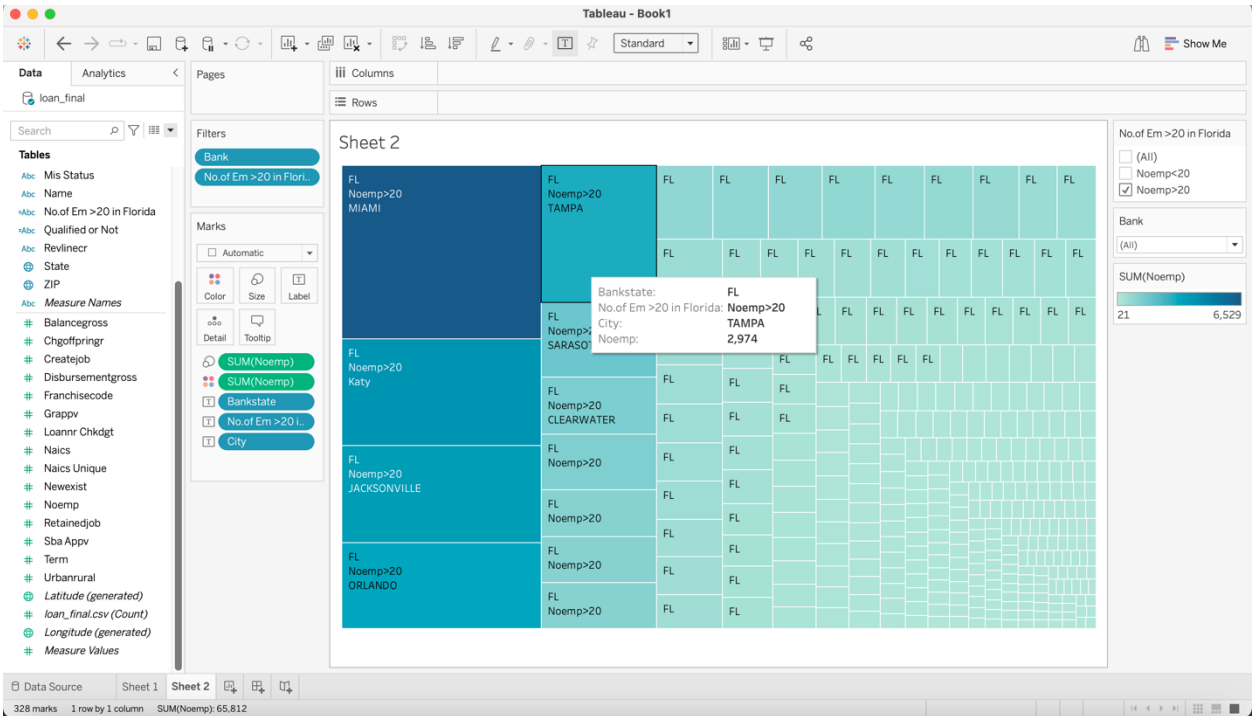
Part 4: Other Topics

Section 4.4: Data Visualization:

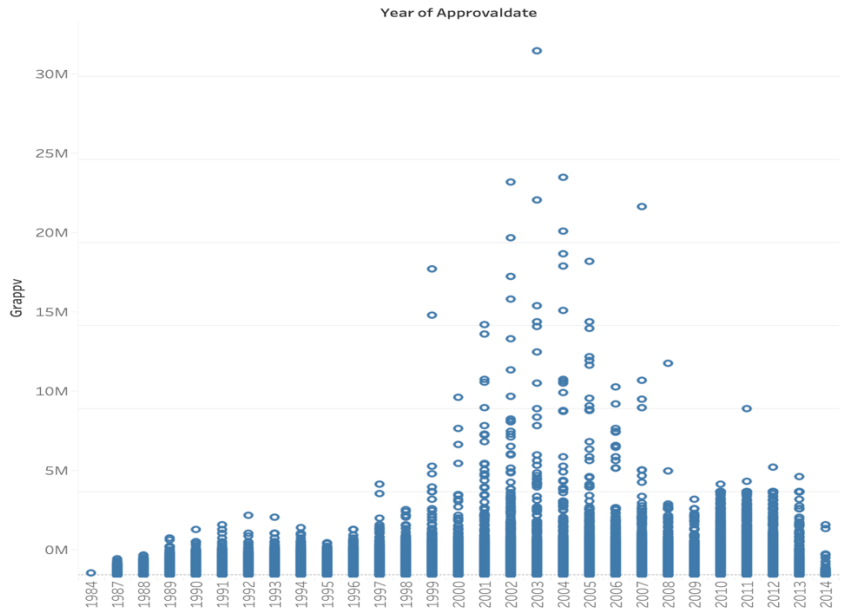
Tableau has been used to visualize data related to the SQL query mentioned in the above sections i.e., the number of businesses who have a loan term greater than 36 months and taken loan after 07/31/2001 are visualized.



The below visualization portrays the businesses with number of employees more than 20 within the state of Florida and segmented by each city in the state.



The below visualization illustrates the comparison of bank-approved amounts and SBA-approved amounts for each company in such a manner that the SBA provides an assurance of approved amounts to the bank, thereby allowing the lender to complete the loan in full. It is possible for the lender to easily determine the amount of funding that has been approved for their company through this visualization.



Lessons learnt

- We have learnt database design and its implementation in a practical way.

References

- Journal of Statistics Education - “Should This Loan be Approved or Denied?”: A Large Dataset with Class Assignment Guidelines - Min Li, Amy Mickel & Stanley Taylor