Assignment 3 – SQL EXCEL integration

# Overview

This week we are going to set up the connectivity between your MySQL databases and Excel, and we will focus on analyzing a loan dataset. The data is sampled from a major financial institution where their data science teams are tasked with developing and comparing several machine learning models to predict “loan status”. Specifically which loans are likely to default.

## Preparation Steps - MYSQL

1. Create a schema assignment\_3, then **use** it.

drop schema if exists assignment\_3;

create schema assignment\_3;

use assignment\_3;

1. Load the SQL query file ***create\_table\_loans.sql*** into your workbench, run through the queries to create table “loan” and table “test\_data”.
2. Check the number of records in the tables. There should be 116283 records in the loan table and 50 records in the test\_data table.
3. DESCRIBE the loan table and check fields and respective data types.

A screenshot of a computer

Description automatically generated

***Data Dictionary:***

* id: A unique identifier for the loan.
* bad\_loan: Indicates whether the loan is bad (1) or not (0).
* loan\_amnt: The amount of the loan.
* int\_rate: Interest rate on the loan.
* emp\_length: Employment length of the borrower in years.
* home\_ownership: The home ownership status of the borrower.
* annual\_inc: Annual income of the borrower.
* purpose: The purpose for which the loan was taken.
* addr\_state: The state in which the borrower resides.
* dti: Debt-to-income ratio.
* delinq\_2yrs: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
* revol\_util: Revolving line utilization rate, or the amount of credit the borrower is using relative to all available credit.
* total\_acc: The total number of credit lines currently in the borrower's credit file.
* longest\_credit\_length: The longest period (in years) over which the borrower has held a credit account.
* verification\_status: Indicates whether the borrower's income was verified.
* term\_length: The term of the loan in months.
* pred: model predicted loan status, bad (1) or not (0) (only in the test\_data table)

## Preparation Steps - EXCEL

1. Review the tutorial slides to set up **ODBC connectivity** on your local computer.
2. Test your connection by running some simple queries here: *Data-Get Data-From Other Sources-From ODBC*

A screenshot of a computer

Description automatically generated

## What to turn in!

Please create an **Excel workbook** and submit that to me via Canvas, for homework assignment credit. Please include your name as part of your file name (e.g., Mike\_Ames\_Assignment\_3.xlsx). Review the tutorial slides to set up ODBC connectivity and write SQL queries in Excel to pull results to different worksheets. In your submission, **all of your queries have to be found in the Excel tab: *Data – Queries & Connections***. I’ve provided example results for the example question to help guide you through this process.

## Results & analysis

### Result 0 – Top 5 purposes of verified loan

**Example Question**: Use the loan table, find the top 5 purposes of verified loans using ODBC. In this and all the following questions, you would need to import data from ODBC: ***Data-Get Data-From Other Sources-From ODBC****.* Then, write a query to extract and load results. Your SQL should produce something that looks like this:

A screenshot of a computer

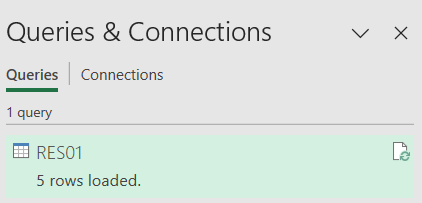
Description automatically generated A screenshot of a computer

Description automatically generated



### Result 1 – TOP default rates by purpose

Write a SQL query to pull data from the loan table, keeping verified loans only. Your query should help to find the top 5 purposes by ranking their average ***default rate (i.e., total # of bad loans / total # of loans)*** indescending order. Load the result set to a worksheet named ‘RES01’ using ODBC and rename your query as ‘RES01’.

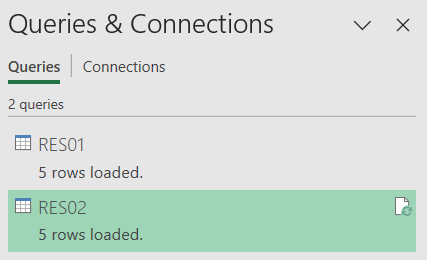


Your SQL should produce something that looks like this:



### Result 2 – TOP default rates by state

Write a SQL query to pull data from the loan table, keeping verified loans only. Among states with more than 1000 loans, find the top 5 states with the highest default rates. Load the result set to another worksheet named ‘RES02’ using ODBC and rename your query as ‘RES02’.

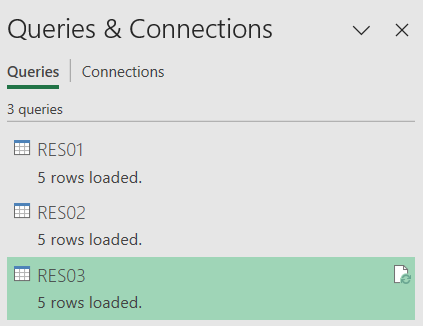


Your result should look something like this.



### Result 3 – average income and loan amount by purpose

Write a SQL query to pull data from the loan table, keeping verified loans only. Your query should help to find the top 5 purposes by ranking their average loan amount in descending order. Calculate average annual income and average default rate for each purpose. Load the result to another worksheet named ‘RES03’ using ODBC and rename your query as ‘RES03’. Your result should look something like this.





Create a **Bar Chart** and a **Lollipop Chart** to visualize the average income and default rate for each loan purpose.

*One example - feel free to design your own charts based on the principles you learned.*

A graph of a graph showing a number of people

Description automatically generated with medium confidence

### Result 4 - confusion matrix

Your colleagues are testing a new default prediction model over a small dataset “test\_data”. Extract appropriate data from test\_data table using ODBC and create a confusion matrix using a ***pivot table*** to evaluate the performance of this classification model.

Calculate 3 evaluation metrics: accuracy, precision, recall. Make sure to explain the meaning of each metric in this business setting and the reason why one should care. Load the result to a worksheet named ‘RES04’ using ODBC and rename your query as ‘RES04’.

*One example, yours can be different:*



### Result 5 – your analysis

Think of a specific question you want to answer using the data. Write a SQL query to obtain the relevant data that answers your question. Load the result to a worksheet named ‘RES05’ using ODBC and rename your query as ‘RES05’. Your question should be interesting and meaningful, your results should be a table. On the same worksheet, create at least one appropriate visualization with 1-2 sentences to describe your findings.

## Rubric

*Out of a total of 25 points*

* *5 questions, 5 points each*
* *All your queries have to be found in the Excel tab:* ***Data – Queries & Connections***

*Your workbook should be well organized. For each question, display your results on a separate worksheet. Your workbook should be clean and organized like this: (Failure to do so may result in a 2-point deduction)*

**