

Lab Work 4: Extraction, Cleansing, Loading and Transformation of Excel Data from Website in Local SQL Server in Monthly Basis

Objective:

This lab aims to implement an ETL pipeline for extracting, cleaning, and loading monthly banking statistics from Excel into a SQL Server database, followed by data validation using SQL queries in SSMS. The objectives include ensuring data accuracy and establishing a reliable process for monthly financial reporting.

Steps:

1. Download the Excel File

- The provided URL was visited and the Excel file was manually downloaded for the latest monthly statistics.

<https://www.nrb.org.np/category/monthly-statistics/?department=bfr>

The screenshot shows the official website of the Nepal Rastra Bank. At the top, there is a navigation bar with links for 'ABOUT', 'LAWS, POLICIES & GUIDELINES', 'REGULATIONS & SUPERVISIONS', 'PUBLICATIONS & STATISTICS', and 'PROCUREMENT'. On the left side, there is a sidebar with links for 'Miscellaneous Notices', 'Directives', 'Circulars', 'Guidelines', and 'List Of BFIs'. The main content area is titled 'Monthly Statistics' and shows a download link for '2082-02(Mid June, 2025) (pdf / xlsx)'. A download progress bar at the top right indicates 'Jestha_2082_Publish-2 (1).xlsx' is 847 KB and Done. There is also a checkbox for 'Don't show when downloads finish'.

2. Install Required Python Libraries

- The following libraries were installed in the python environment:

```
pip install pandas sqlalchemy pyodbc openpyxl
```

```
PS D:\Aarchi_022bim003> pip install pandas sqlalchemy pyodbc openpyxl
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: pandas in c:\users\dell's\appdata\local\packages\pythonsoftwarefoundation.python.3.13\local-packages\python313\site-packages (2.3.0)
Collecting sqlalchemy
  Downloading sqlalchemy-2.0.41-cp313-cp313-win_amd64.whl.metadata (9.8 kB)
Collecting pyodbc
  Downloading pyodbc-5.2.0-cp313-cp313-win_amd64.whl.metadata (2.8 kB)
Collecting openpyxl
  Downloading openpyxl-3.1.5-py2.py3-none-any.whl.metadata (2.5 kB)
```

3. Extract the Excel file and Perform Data Cleansing

- The C14 Sheet of the 'Jestha_2082_Publish-2.xlsx' file was extracted into a new 'Monthly_statistics.xlsx' file.

A screenshot of Microsoft Excel showing the 'Districtwise List of Banks and Financial Institutions Branches (Jeth, 2082)' on the C14 sheet. The table has columns for S.No., Province, District, Class 'A', Class 'B', Class 'C', Total, Population, and Pop. Per Branch. The data includes rows for Koshi districts like Taplejung, Panchthar, Ilam, Jhapa, and Sankhuwasabha, with their respective branch counts and populations.

S.No.	Province	District	Class 'A'	Class 'B'	Class 'C'	Total	Population	Pop. Per Branch
1	Koshi	Taplejung	24	-	-	24	120,590	5,025
2	Koshi	Panchthar	28	3	2	33	172,400	5,224
3	Koshi	Ilam	38	9	1	48	279,534	5,824
4	Koshi	Jhapa	139	54	12	205	998,054	4,869
5	Koshi	Sankhuwasabha	32	1	-	33	158,041	4,789

- The data cleaning process was performed by first removing the redundant 'S.No.' column and eliminating the summary 'Total' rows to focus solely on district-level data. The column names were standardized by replacing special characters and whitespaces with underscores (e.g., 'Class 'A'' was converted to 'class_a') to ensure SQL compatibility. Additionally, the dataset was enriched by adding a 'month_year' column containing 'Jestha, 2082' for each record, enabling proper temporal tracking of the banking statistics.

A screenshot of Microsoft Excel showing the cleaned 'Monthly_statistics' data on the K81 sheet. The table has columns for Province, District, Class_A, Class_B, Class_C, Total, Population, Pop_Per_Branch, and month_year. The data is identical to the original table but lacks the redundant S.No. and Total columns, and includes the month_year column.

	A	B	C	D	E	F	G	H	I
1	Province	District	Class_A	Class_B	Class_C	Total	Population	Pop_Per_Branch	month_year
2	Koshi	Taplejung	24	-	-	24	120,590	5,025	Jestha, 2082
3	Koshi	Panchthar	28	3	2	33	172,400	5,224	Jestha, 2082
4	Koshi	Ilam	38	9	1	48	279,534	5,824	Jestha, 2082
5	Koshi	Jhapa	139	54	12	205	998,054	4,869	Jestha, 2082
6	Koshi	Sankhuwasabha	32	1	-	33	158,041	4,789	Jestha, 2082

4. Code Execution & Database Upload Process

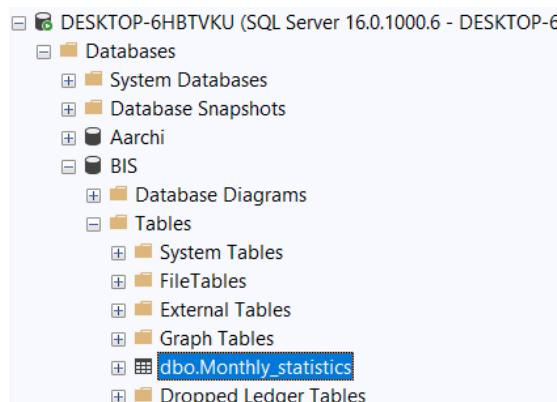
- Script Execution:** The Python script connects to the BIS database in SQL Server using SQLAlchemy and pyodbc, then processes all Excel/CSV files in the specified directory. It reads each file into a Pandas DataFrame, cleanses the data by removing commas from numerical values, and prepares it for database insertion.
- Database Loading:** Using df.to_sql(), the script dynamically uploads each file's data to a corresponding table in SQL Server (named after the source file). The if_exists='append' parameter ensures monthly data accumulates in the same table, while the script logs successful uploads for verification.

```

PS D:\Aarchi_022bim003\Lab_4> python -u "d:\Aarchi_022bim003\Lab_4\Lab_4.py"
=====
CONNECTION SUCCESSFUL
=====
Importing file: D:\Aarchi_022bim003\Lab_4\Monthly_statistics.xlsx
Number of records: 77
First few rows:
   Province      District Class_A Class_B Class_C Total Population Pop_Per_Branch month_year
0   Koshi     Taplejung      24       0       0    24    120590  5024.583333333333  Jestha 2082
1   Koshi    Panchthar      28       3       2    33    172400  5224.242424242424  Jestha 2082
2   Koshi        Ilam      38       9       1    48    279534           5823.625  Jestha 2082
3   Koshi      Jhapa     139      54      12   205    998054  4868.556097560976  Jestha 2082
4   Koshi  Sankhuwasabha      32       1       0    33    158041  4789.121212121212  Jestha 2082
File Monthly_statistics.xlsx imported successfully into table 'Monthly_statistics'.
=====
FILES IMPORTED SUCCESSFULLY
=====
```

5. Verify Database Upload

- The successful upload of the cleaned data to the monthly_statistics table in the ‘BIS’ database in Microsoft SQL Server Management Studio (SSMS) was verified through SQL query validation after executing the ETL process, ensuring data integrity and completeness.



- The following image shows the successfully executed result of the SQL query mentioned above Microsoft SQL Server Management Studio (SSMS), displaying data from the monthly_statistics table.

The screenshot shows a SQL Server Management Studio window titled "SQLQuery1.sq...Dell's (73)*". A single query is present: "SELECT * FROM Monthly_statistics;". The results grid displays data for 7 rows, each representing a district in Koshi Province. The columns are Province, District, Class_A, Class_B, Class_C, Total, Population, Pop_Per_Branch, and month_year. The data includes various districts like Taplejung, Panchthar, Ilam, Jhapa, Sankhuwasabha, Bhojpur, and Terhathum, along with their respective statistics and population density.

	Province	District	Class_A	Class_B	Class_C	Total	Population	Pop_Per_Branch	month_year
1	Koshi	Taplejung	24	0	0	24	120590	5024.583333333333	Jestha 2082
2	Koshi	Panchthar	28	3	2	33	172400	5224.242424242424	Jestha 2082
3	Koshi	Ilam	38	9	1	48	279534	5823.625	Jestha 2082
4	Koshi	Jhapa	139	54	12	205	998054	4868.556097560976	Jestha 2082
5	Koshi	Sankhuwasabha	32	1	0	33	158041	4789.121212121212	Jestha 2082
6	Koshi	Bhojpur	24	0	1	25	157923	6316.92	Jestha 2082
7	Koshi	Terhathum	21	3	0	24	88731	3697.125	Jestha 2082

Conclusion:

The key takeaways from this lab include the following important aspects:

1. **Data Integration:** Extract, clean, and upload Excel data into SQL Server efficiently using Python.
2. **Automation & Scheduling:** Automate monthly data import tasks with dynamic scripting and scheduling tools.
3. **ETL Proficiency:** Develop foundational ETL skills for structured data handling, storage, and analysis.