Overview

This document provides an information about what I tried to do at the technical and data analytic level based on given data. Firstly, I tried to build ETL data pipeline steps to store the data in SQL Server from JSON. Because, so far, the SQL Server is the tool I have only to store the data on my local machine.

Also, at the data visualization level, I tried to demonstrate my Python skill to build some dashboards for some potential questions.

Technologies:

* Data Ingestion side: Python (Pandas, NumPy)
* Data storage side: SQL Server DB
  + SQL Queries for data aggregation side from SQL Server to Python
* Data Visualization side: Python (Pandas, Plotly, Matplotlib)

Tools:

* PyCharm (for data ingestion in Python)
* Jupyter Notebook (for data visualization)
* SSMS (SQL Server management studio)

Data pipeline structure

In terms of the data ingestion and storage level, I developed 3 phase of python codes from the source side to the target dashboard side.

|  |  |  |  |
| --- | --- | --- | --- |
| **Source data** | **Data Ingestion** | **Data Warehouse** | **Data Visualization** |
| Json file document icon Royalty Free Vector Image   * Loan Data * Pool option Data * Eligible price combination * Baseline | Why Anaconda's Data Science Tent Is So Big--And Getting Bigger  Data Loader from JSON to SQL Server in Python   * LoanDataLoad.py | Microsoft Sql Server Icon PNG Transparent Background, Free Download #11376  - FreeIconsPNG  Local database in my Laptop   * DB Name: MORTGAGEDATA * Server: Local DB Server | Plotting a stress-strain curve with four libraries: matplotlib, pandas,  altair and bokeh - Python for Undergraduate Engineers   * Read aggregated data from SQL Server * In Python using matplotlib library |
| * A reason I used Python for the data ingestion and data visualization side is that I wanted to demonstrate my python skills for both of these 2 libraries. * Likewise, I could implement dashboard directly from JSON in Python. However, as a ETL pipeline perspective, I wanted to demonstrate my SQL query skills with python even though there were not too much complex queries needed to developed. | | | |

Data analysis

Based on my investigation, I the following questions to be asking for the pooling problem for given data.

1. As we can see from the following dashboard, there are 11 pools having not so good scored loans. Based on that, would it be suitable to involve these following pools?

|  |
| --- |
|  |

I would like to suggest you/your company don’t have to include these following 11 pools into the repeating cycle.

|  |
| --- |
|  |

1. How many pools with high balanced loans more than 10%? As we can see from the picture below, there are 32 pools and 16 ones of them are part of Fannie Mae and rest of 16 ones are Freddie Mac.

|  |
| --- |
|  |

The following list and the file named “Full list.xlsx” shows us detailed list of those 32 pools that we can choose.

|  |
| --- |
|  |

Files description

|  |  |
| --- | --- |
| File name | Description |
| LoanDataLoad.py | Data loader from JSON to SQL Server |
| Question1.py | Data visualization based on SQL Query joins |
| Question2.py | Data visualization based on SQL Query joins |
| Full list.xlsx | List of pools that I’m suggesting to take |
|  |  |