BNZ Advanced Analytics MLE Technical Exercise

This is an open-ended exercise and opportunity for you to demonstrate your capabilities and skills with data. Use the following StatsNZ datasets (https://www.stats.govt.nz/assets/Uploads/Business-employment-data/Business-employment-data-March-2024-quarter/Download-data/business-employment-data-march-2024-quarter.zip and https://www.stats.govt.nz/assets/Uploads/Business-financial-data/Business-financial-data-March-2024-quarter/Download-data/business-financial-data-march-2024.zip) to answer questions 1 to 5.

We request the answers be written in a Jupyter notebook or Google Colab notebook (see https://colab.research.google.com/#scrollTo=OwuxHmxllTwN) and then committed to a Github repository. Once you have completed the exercise, please email a link to your repository.

We would like to see the code you used to obtain the answers to questions 1 to 3 and **request that you do so using SQL operations performed using DuckDB.** See https://duckdb.org/ for documentation on how to do this.

Provide the answer and write a DuckDB SQL query that produces it.

For question 5 you are free to use the methodology of your choice (we have a preference that you use Python if possible).

1. Of the industries where salaries and wages data did NOT exist in 2016 and only appeared later, which industry had the highest average value for actual filled jobs across time and what was that value?

2. Provide the answer and write a DuckDB SQL query to show which year/month combination and industry had the second highest seasonally adjusted operating income sales across all the business industries in New Zealand that are categorised as NZSIOC level 2.

3. Create a DuckDB SQL query to calculate the quarterly cumulative number of filled jobs over time for the territorial authority with the highest average value of filled jobs across time. **You may not use window functions in your query.**

4. Assume these datasets are used in part of a pipeline where the file that arrives may contain unwanted duplicates, incorrect datatypes, missing dates or other data quality aberrations. What things could be done programmatically to make sure the input data is of adequate quality and improve the pipeline?

5. Create summary statistics and perform a statistical analysis or create a model using the provided datasets. We are interested in your justification for your choices and reasoning. You may join the data to other datasets from https://www.stats.govt.nz/large-datasets/csv-files-for-download/ if you wish.