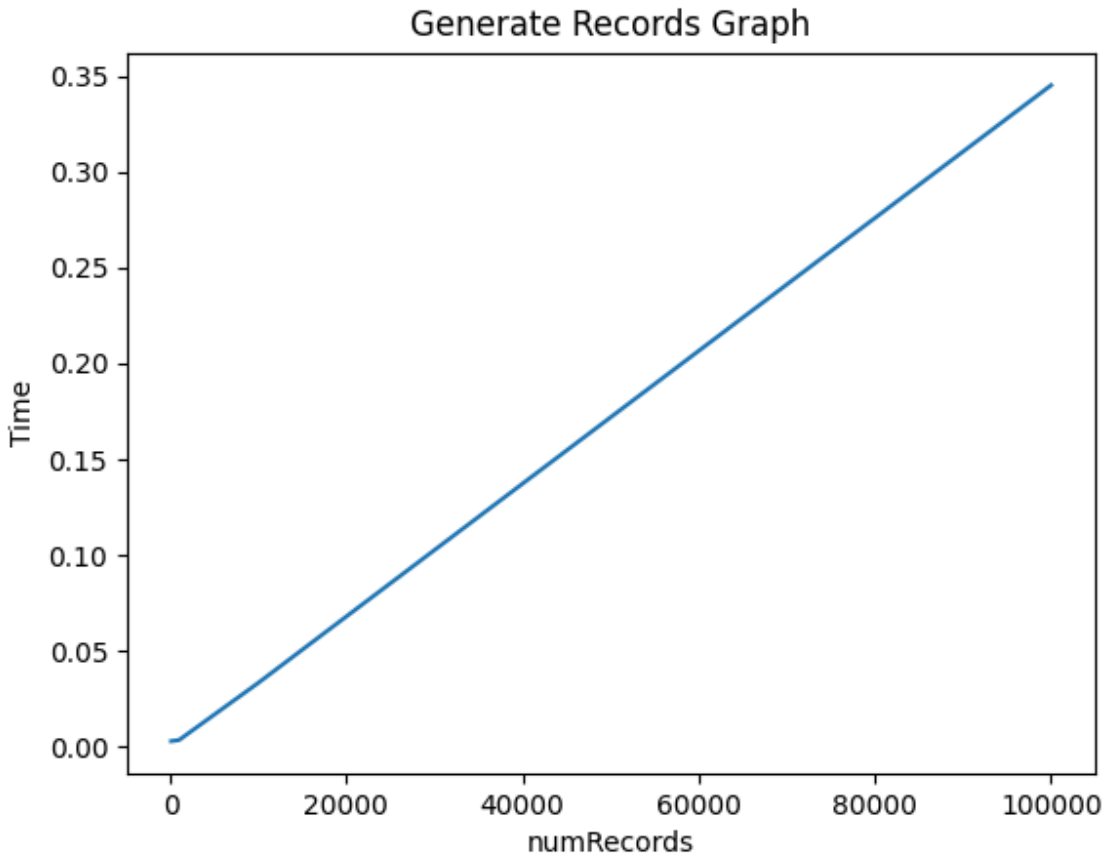
Name: Ganesh Kumar
Date: 02/17/2025
Instructor: Professor Andres Calle
Course: CIST-005B-34571 Advanced Python
Performance Trends:
How does each operation's execution time change as the dataset grows?

- Load in sales data (reading from a CSV or database):
  The steps to load in sales data are:
  - Open SalesRecords.csv file in reading mode.
  - while not end of file, read entries one line at a time.
  - Split and extract the sale id, date, price, and product.
  - Create a dictionary entry for salesRecords and add the salesRecords.
  - if end of file, exit.
  - Time complexity is O(n).
- Retrieve the latest sale:
  The steps to retrieve the latest sale are:
  - Loop through the salesRecords dictionary, reading one record at a time from the dictionary.
  - Split and extract the year, month, and day.
  - Compare the year, month, and day with the latest year, latest month, and latest day.
  - Update the latest year, month, and day.
  - Until all records are processed.
  - The time complexity is O(n).
- Compute the total revenue:
  The steps to compute the total revenue are:
  - Loop through the dictionary numRecords times.
  - Retrieve each record.
  - Add the sale item price to the total revenue.
  - The time complexity is O(n).
- Check for duplicate sale IDs:
  - Create a duplicateSales list.
  - Read each sale record from the SalesRecords.csv file.
  - Search the id number in salesRecords dictionary.
  - If the item is duplicate,

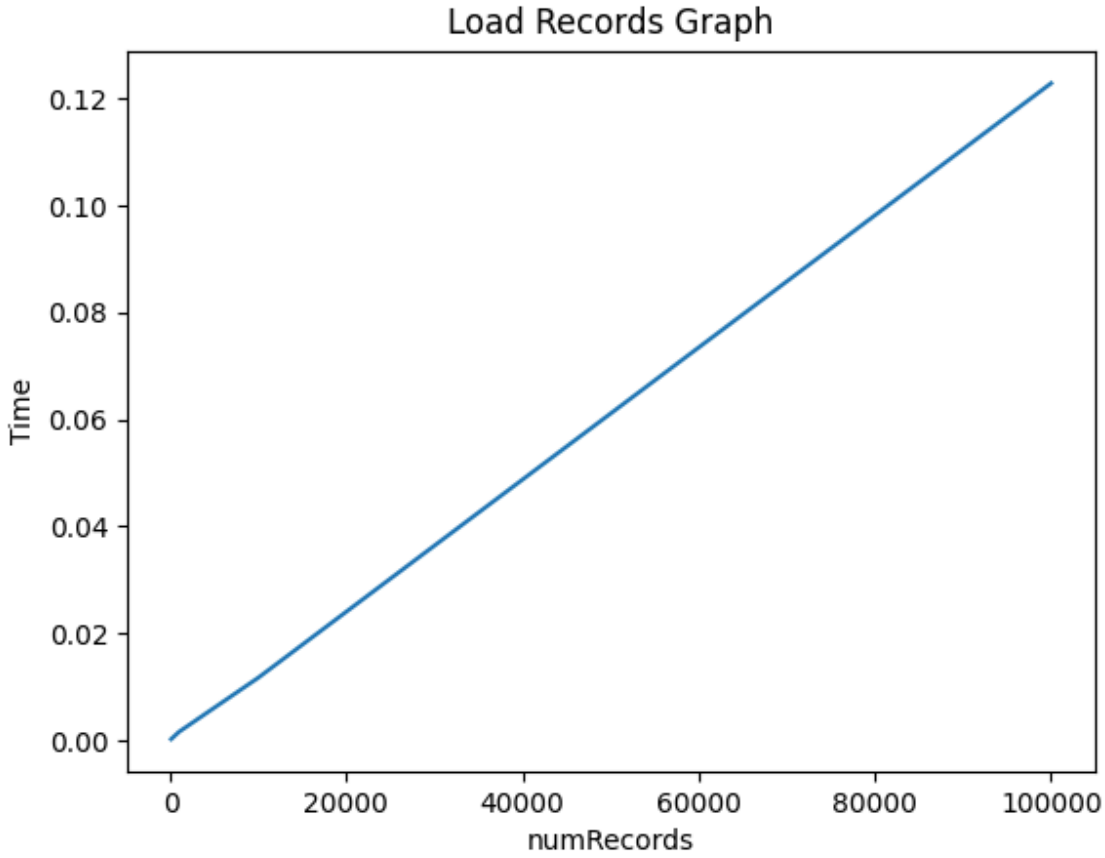
    - Add it to the duplicateSales list.
    - The time complexity is $O(n^2)$ because every item is checked numRecords times, resulting in the time complexity of $O(n^2)$.
- Search for a sale by its ID:
    - To search a random sale ID in the dictionary.
    - Compare the sale ID with each entry in the dictionary.
    - If a match is found:
        - Print the sale record.
    - Else:
        - Print that it is not found.
    - Time complexity is O(n) because you need to search all the entries once, even if the entry is not in the dictionary.

Do the results align with the theoretical Big O expectations?

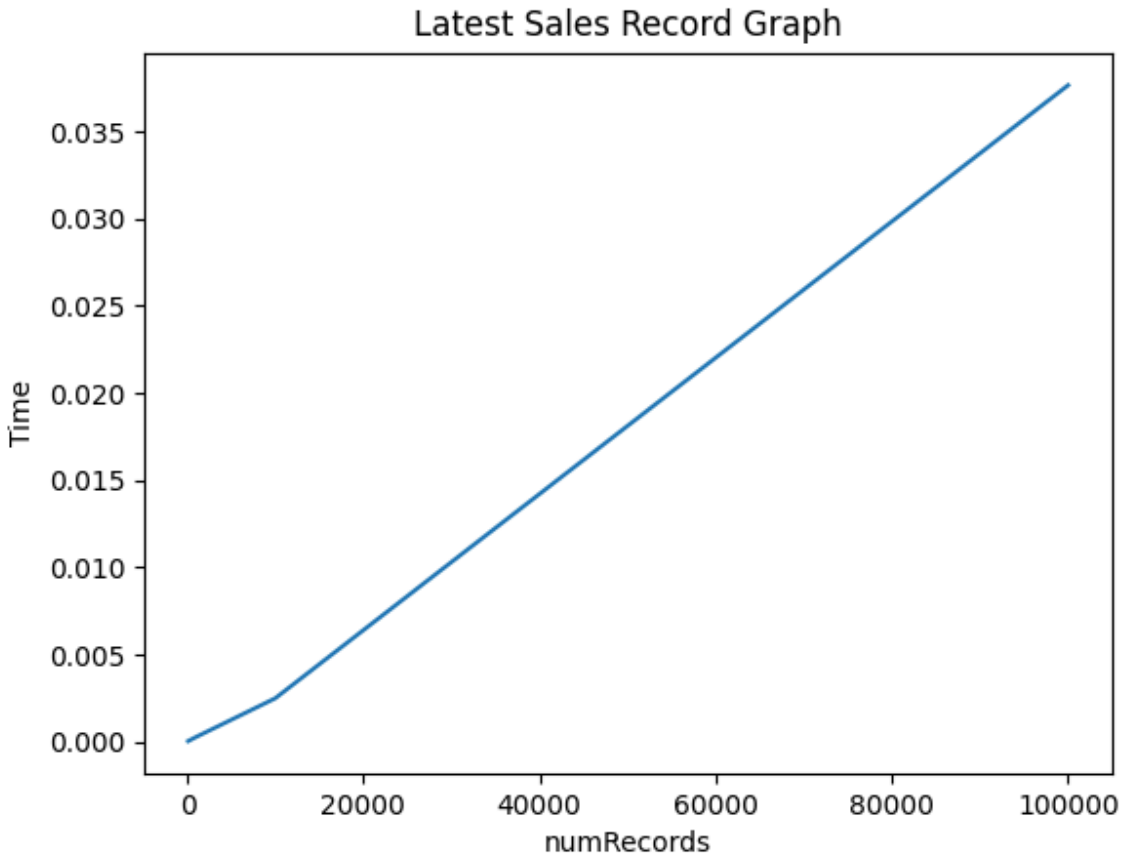
The time complexity graphs below show the number of records on the x-axis and the time on the y-axis for different operations.



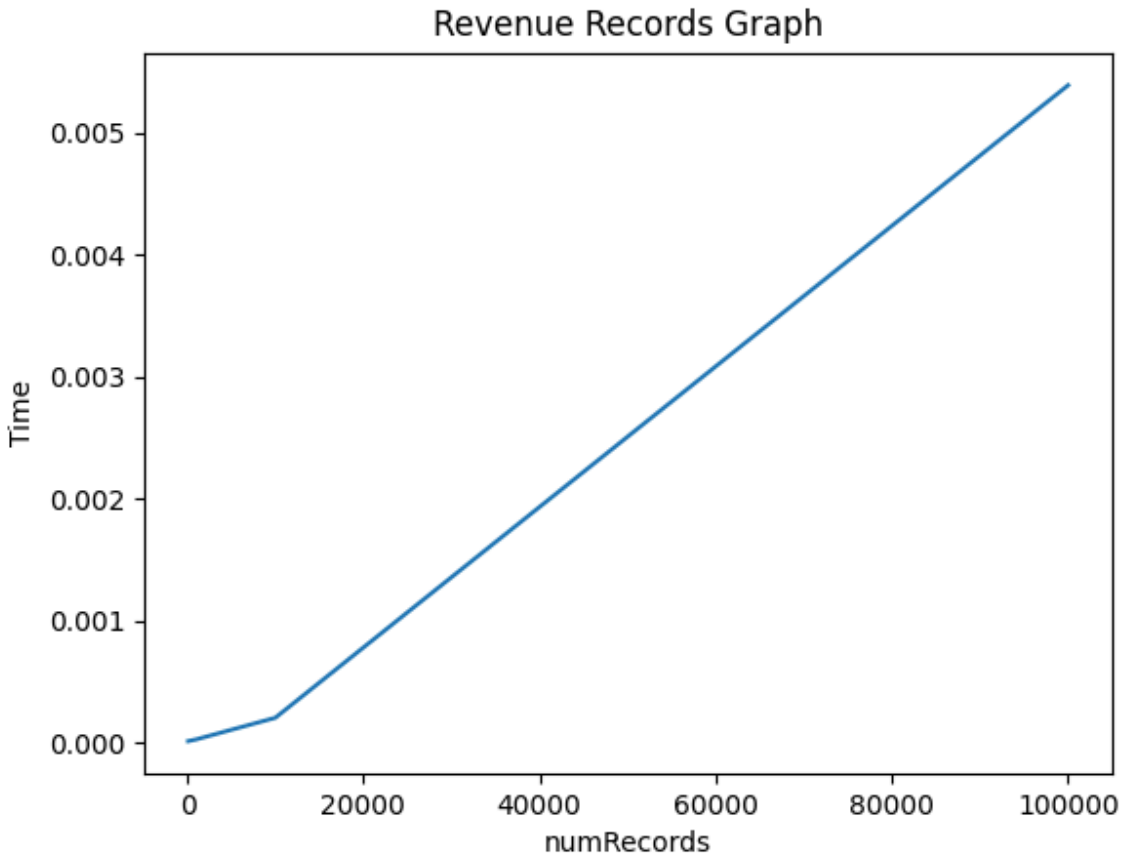
The results agree with Big O notation because generating numRecords is a linear algorithm with time complexity of O(n).



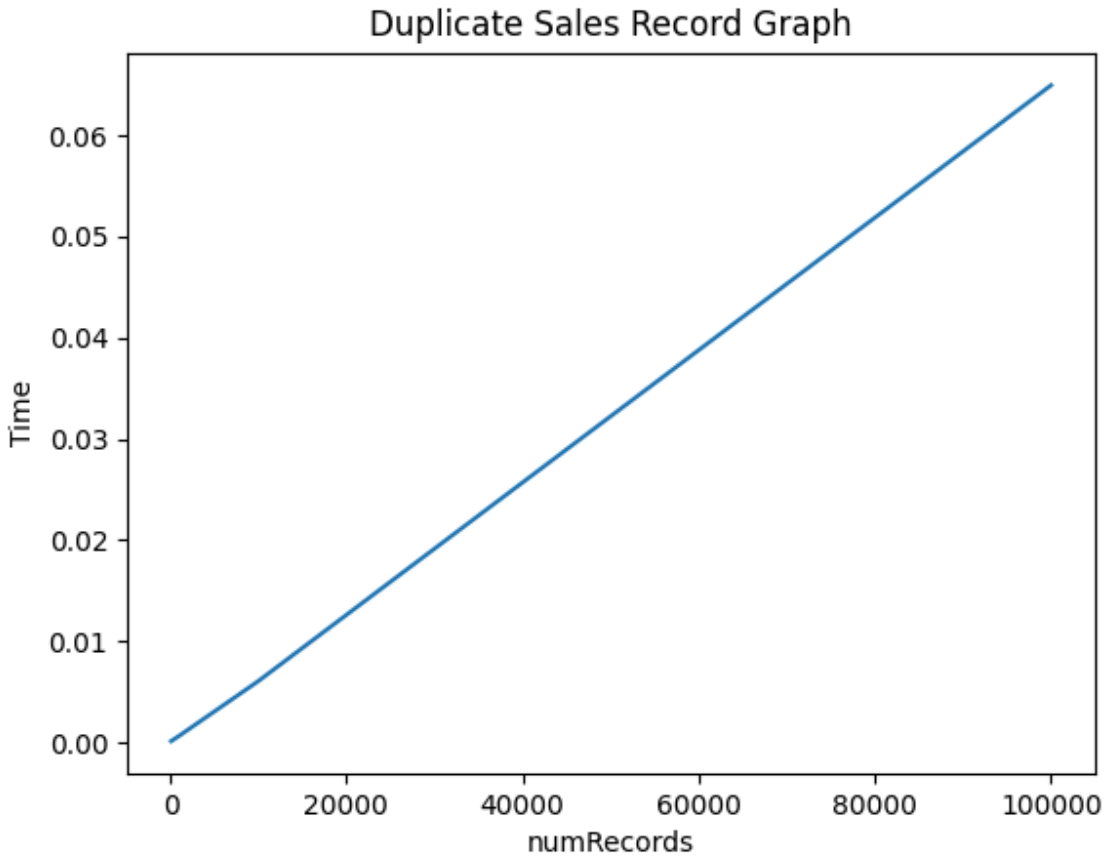
The results agree with Big O notation because loading numRecords is a linear algorithm with time complexity of O(n).



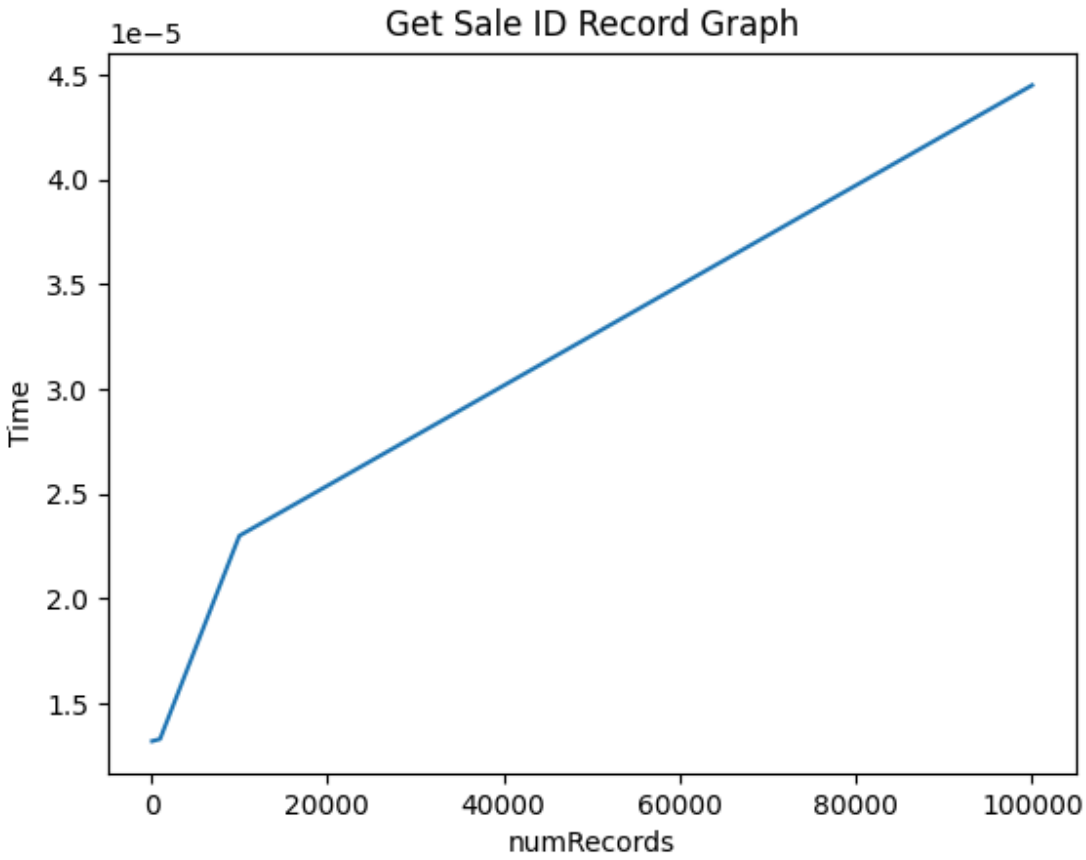
The results agree with Big O notation because finding the latestSales is a linear algorithm with time complexity of O(n).



The results agree with Big O notation because calculating the revenue is a linear algorithm with time complexity of O(n).



The results disagree with Big O notation because finding the duplicate sales is a quadratic algorithm with time complexity of $O(n^2)$.



The results generally agree with Big O notation because finding the duplicate sales is a linear algorithm with time complexity of O(n). But it is not clear why the slope is different from 0 to 10,000 records compared with the slope from 10,000 to 100,000.

Real-World Implications:

Which steps might become bottlenecks in a production system processing millions of records?

Finding the duplicates might become bottleneck because finding duplicates takes $O(n^2)$, so quadratic algorithms are inefficient.

How would you optimize or replace the inefficient (quadratic) approach?

I would improve it by sorting the dictionary using heap sort algorithm, which has the time complexity of O(n log n). When locating a duplicate entry, only one comparison is needed to check for duplicate entries.

Practical Adjustments:

How might you put together a testing plan for this project?
Generate sales records with negative sale id and negative prices along with negative number of records and incorrect date format and negative revenue.
What additional error handling or data validation would be necessary?
The additional error handling or data validation that would be necessary is checking the correct number of days in a month or incorrect sale id numbers