Python is becoming a top language as well as a top priority for data analysis because of its ease of use, adaptability, flexibility, and robust and powerful libraries. I’ve learned several things about using Python as programming language for data analysis from my experience which I will share in detail below.

Developed Jupyter notebook to analyse data using simple python testing the code’s usability with various datasets and comparing its performance to specialised libraries like NumPy and Pandas. I also investigated the usefulness of utilising ChatGPT and other AI-driven solutions to help with these chores.

Dataset used for each task is attached in the appendix.

**Learnings and Insights**

**Importing Data**: - The first step in data analysis is importing data into python. Because of robust libraries Python offers several techniques of importing data from reading a CSV, Excel files and databases using strong built-in modules & libraries like csv and pandas. I used both methods in task 1 and task 2 respectively gaining a hands-on experience in importing, reading, and processing csv files using python.

I found that importing a file with pandas is much more user friendly and convenient as compared to csv because importing data with csv includes lots of steps to uncover before we start working with the data on the other hand pandas is just a one step process after which our data is ready to be played upon.

The key difference between both is that in csv the data is typically read into a sublist of list and accessing te data requires manual indexing where as with pandas the data is read into a data frame (a data structure especially designed for data analysis) making it easier for accessing and manipulating data.

**Data Cleaning**: - Addressing missing values, eliminating duplications, and fixing data types are all part of Data analysis or Data cleaning (to be accurate) in Python. These tasks can be easily performed by rich libraries like NumPy and pandas but can also be performed with the help of basic fundamentals python concept like loops, conditional operators (if else elif), list manipulations, Slicing, Inbuild functions such as .append() , type casting and many more.

I used all the above-mentioned fundamental python concept in Task 1 for data cleaning and concluded that importing/using Python libraries make it easier for analysing data as it is more efficient, less time consuming, taking memory usage into consideration.

Data cleaning tasks in pure python require writing complex code and involve manually handle operations like conditional checks, iterations , looping through data and much more but pandas does the same task in few or single line of code.

**Performance Testing**: - Used Python’s time module to measure and compare the execution times of various code sections or code blocks.

**Calculating Statistics**: - Implemented and created functions with loops to calculate the summary statistics of data which include mean, median, mode, standard deviation and varience of the dataset. I have also calculated unique values present in a column, their count, their range and the minimum and maximum values present in a column.

The above calculations were achieved with the help of fundamental concepts of python in Task 1 utilizing everything and making a single code block for it and in Task 2 these calculations were done with the liberty of pandas making it hassle free and easier and more convenient with just a single line of code.

It can be observed that utilizing pure python for analysis of huge datasets has its own limitations especially when it comes to memory utilization and execution time. On the other hand, pandas being based on C it can handle large datasets more effectively for analysis.

**Task 3** was all about testing the usability of code developed in task 1 on a random csv file from the web such that there are minimal modifications to adapt to the dataset.

I discovered that in order to accommodate various data structures, datatype and formats the code needed to be modified in multiple ways including changing column indices , slicing issues, fixing column datatype, issues with handling missing values and data cleaning procedure as well.

From my overall experience I learned that though fundamental python approach is flexible and adaptable, but it is not as reliable or convenient as employing data analysis packages and robust libraries.

And also the requirement to make code modification indicates that the fundamental python usage is impractical in real world analysis.

**Part 2 task 3**

The prompts I used are attached in the appendix which shows how smart and clear AI is in understanding the problem statement and giving the desired output. Noted that while chatgpt is very powerful and useful it requires correct and accurate prompting and validation to generate the desired code.AI tools can increase the coding process significantly. It gives useful code snippets, debug our code as well as offers alternate suggestions and explanations.

**Conclusion: -**

Using Python for data analysis has been a satisfying experience that has improved my comprehension of both the basic idea of programming and the real-world difficulties in managing the data. For effective analysis optimised libraries like NumPy and Pandas are essential even though python fundamental provide strong base to all.

Appendix : -

Dataset for task 1:-

<https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction/data?select=train.csv>

Dataset for task 2:-

<https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction/data?select=test.csv> and

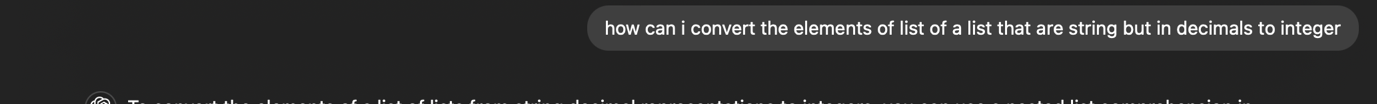
<https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction/data?select=train.csv> were combined to increase the size of the dataset approx. 7000 rows were added

Prompts asked to chatgpt to debug and assist me with coding:-

A black background with white text

Description automatically generated

Can be helped to understand a wide variety of things from little minute things to huge code.



A black and white sign with white text

Description automatically generated

Used some functions which were not adviced to use hence telling chatgpt to not use them.

A screenshot of a computer

Description automatically generated

Gave the solution for list comprehension but it helped me understand the logic. It shows giving the accurate prompt is really necessary to get the desired result.

A screenshot of a computer

Description automatically generated

Using it for debugging my code.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated