Do you find one approach, easier or more performant?

I find that Pandas is often the easiest approach for most people because it's well-documented, widely used, and offers a lot of flexibility. It integrates well with various other tools and has a large community for support. However, when working with larger datasets or trying to optimize for speed, Polars is more performant. It's built to handle large volumes of data in a more efficient way, especially when you're dealing with operations that would otherwise be slower in Pandas. For simpler tasks or smaller datasets, Pandas is a bit more straightforward, but for handling big data, Polars shines. In terms of performance, Polars has a definite edge, particularly when working with multi-threaded environments.

If you were coaching a junior data analyst, what approach would you recommend?

If I were coaching a junior data analyst, I’d recommend starting with Pandas. It’s the industry standard, and there are tons of resources available to learn from. Plus, it’s easy to get up and run with, especially for those new to data analysis. As they grow in their understanding of data processing and performance optimization, I would introduce them to Polars, especially for larger datasets or when they encounter performance bottlenecks in Pandas. Starting with Pandas allows them to learn the basics and the logic of data manipulation, and then they can progress to more advanced tools like Polars as they get more experienced.

Was it a challenge to produce identical results?

It wasn't a major challenge, but there were definitely a few nuances that required attention. The core logic behind the calculations is the same across all approaches, but each tool (Pure Python, Pandas, and Polars) has its own way of handling missing values, data types, and performance optimizations. For instance, Pandas and Polars both provide describe (), but the way they handle non-numeric columns or missing values differs slightly, which requires a bit of fine-tuning to ensure the outputs matched.

In Pure Python, the challenge was more about manually implementing certain functions like standard deviation, which is already built-in with Pandas and Polars. While the results were consistent, I had to be careful with the data types and how each tool dealt with the data under the hood, especially with nullable types or categorical data.

Overall, the process was a good learning experience, and while there were some minor adjustments to make the results match exactly, it wasn’t too difficult. It’s all about understanding the intricacies of each tool and making sure they handle the same operations in a consistent manner.