Descriptive Statistics Analysis Report

Name:Kulveen Kaur

University:Syracuse University

Course:Applied Data Science

Task:Research Task 04 - Descriptive Statistics Using Python, Polars, and Pure Python

# Introduction

The objective of this task was to compute and compare descriptive statistics using three different approaches — Pandas, Polars, and pure Python. The dataset used, 'MERGED2021\_22\_PP.csv', contains educational data for institutions across U.S. states. The goal was to group the data by state abbreviation (STABBR) and perform analysis on selected features such as median earnings, Pell Grant percentages, and graduate debt.

# Tools & Libraries Used

- Python 3  
- Pandas  
- Polars  
- Python built-in data structures  
- Visualizations with Matplotlib

# Methodology

Each script — 'pandas\_stats.py', 'polars\_stats.py', and 'pure\_python\_stats.py' — was designed to:  
1. Read the CSV data.  
2. Group by the STABBR column.  
3. Calculate count, mean, min, max, standard deviation (for numeric columns), and frequency counts (for categorical columns).  
4. Save visualizations in a 'visuals/' folder.

# Performance Comparison

Based on the performance report generated:  
- Polars was the fastest in terms of execution time, efficiently handling large datasets with optimized memory usage.  
- Pandas provided balanced performance and ease of use.  
- Pure Python had the slowest execution time but gave insight into underlying computations.

# Visualizations

Visualizations such as bar charts and histograms were generated to illustrate the distribution of earnings and Pell Grant recipients across states. These are saved in the 'visuals/' folder.

# Conclusion

This task deepened understanding of descriptive statistics and allowed for comparison of performance across different data processing techniques. It demonstrated the trade-offs between ease of use (Pandas), speed (Polars), and control (Pure Python).