Syracuse University
School of Information Studies

OPT BI-MONTHLY ACCOUNTING FORM

We thank you for your time spent taking this survey.
Your response has been recorded.

Name (Last Name, First Name):

Purohit Bhavesh Pratap

Faculty Sponsor:

● **Jeffrey Saltz**
○ Ingrid Erickson

Reporting Period:

○ May 1- 15, 2025

○ May 16 - 31, 2025

○ June 1 - 15, 2025

● **June 16 - 30, 2025**

○ July 1 - 15, 2025

○ July 16 - 30, 2025

○ August 1 - 15, 2025

○ August 16 - 30, 2025

I worked on my assigned OPT project during the past two weeks

● **True**

○ False

I would summarize my activity during this time period as follows:

During this period, I focused on performing descriptive statistical analysis of datasets related to the 2024 US Presidential election. I implemented this analysis using three different programming approaches: Pure Python (Standard Library) – to understand low-level data handling and computation without relying on third-party tools. Pandas – the de-facto data analysis library, used for its ease, flexibility, and rich functionality. Polars – a modern, high-performance DataFrame library optimized for speed and memory efficiency. Each approach was applied to three different datasets (Facebook ads, Facebook posts, Twitter posts), requiring consistent preprocessing to achieve comparable results. Special attention was given to cleaning nested or complex fields, handling missing values, and generating accurate summaries. The outputs were then compared for consistency across methods.

Upon reflection, this project is allowing me to learn or refine the following STEM-related skills:

a. Data Cleaning and Preprocessing: Gained hands-on experience in transforming messy, real-world datasets into analysis-ready formats. Learned techniques to flatten nested JSON-like fields (e.g., demographic_distribution, delivery_by_region) and normalize inconsistent data types and missing values. b. Descriptive Statistics and Aggregation: Improved my understanding of core statistical measures such as count, mean, median, min/max, and unique counts. Practiced grouping and subsetting data to extract deeper insights, particularly using .groupby() in Pandas and Polars. c. Programming Concepts and Tool Comparison: Reinforced my Python programming fundamentals, especially around working with dictionaries, lists, and basic file I/O in the pure Python scripts. Explored the strengths and trade-offs of Pandas vs. Polars, especially in terms of performance, syntax clarity, and flexibility. d. Reproducibility and Standardization: Ensured consistency across implementations by carefully designing reusable processing logic. Gained an appreciation for standardizing workflows, data types, and formatting for reliable analysis. e. Problem Solving and Debugging: Addressed challenges like floating-point precision mismatches and inconsistent null values. Developed strategies to test and validate statistical outputs across multiple tools. f. Technical Communication: Practiced clearly documenting the project's structure, purpose, and results in a detailed and accessible README.md. Reflected on tool recommendations for junior analysts and noted the practical value of AI-assisted code generation.

Additional comments:

This project was both technically challenging and deeply rewarding. Working across three tools for the same analysis helped me understand not only the "how" but also the "why" behind certain library choices in data science workflows. It reinforced the importance of data quality, careful preprocessing, and methodological rigor when working with complex datasets. It also highlighted the growing relevance of high-performance tools like Polars in modern data analysis. Overall, this task helped bridge theoretical statistical knowledge with practical data science implementation skills, which I consider essential for any aspiring analyst or researcher in a STEM field.