## ****Research Task 4:****

## ****Descriptive Statistics with and without 3rd Party Libraries (Pandas/Polars)****

### ****Welcome****

**This period’s task is to build a data summarizing system that will allow a researcher to give a dataset and have a descriptive summary or analysis of that dataset to be constructed for the researchers. To that end you will build (or build upon) a system to produce that analysis. This task includes three datasets all related to the 2024 US presidential elections and social media activity. A small cohort of students may be a step ahead of you having already been exposed to the “ad” dataset. Feel free to discuss their work and findings with them or me if you would like to get a jump start. Remember this research is as much for you as it is for me so, please tackle as much as you can yourself.**

**If you completed this task last period, your task this period will be to extend your statistics system to handle arbitrary datasets, not strictly those from before. To this end, you will likely want to drive your analysis off of some set of meta-data that describes the actual data to be analyzed. Please select a small number of public datasets of your choosing. Kaggle might be a good place to find public data.**

### ****Objective****

Your task is to explore and summarize a real-world dataset using **base Python**, **Pandas,** and **Polars**. With this research task I am looking for you to produce the same numerical results via all three strategies. I’m not only interested in the results, but your approach and feedback on how the three strategies compare. Was it a challenge to produce identical results? If so, please describe the challenge and how you overcame it. Do you find one approach, easier or more performant? If you were coaching a junior data analyst, what approach would you recommend? Can coding AI such as chatGPT produce recommendations such as template code to jump start each approach? What default approach do these kind of tools recommend when asked to produce descriptive statistics? Do you agree with these recommendations (why or why not)? Note that answering some of these questions will require data cleaning as not all columns are simple values, and you will need to unpack some column values into new columns. You might also need to use time measuring counters to determine relative performance. Note that while I assume you are using Python here, there is no reason that R or some other alternative approach could be used. If you do use some alternative, please be sure to approach the task from a base library vs 3-rd party library perspective.

### ****Step 1: Download the Dataset (this is your first assignment)****

<https://drive.google.com/file/d/1Jq0fPb-tq76Ee_RtM58fT0_M3o-JDBwe/view?usp=sharing>

This data set represents Facebook ads, Facebook Posts, and Twitter Posts during the 2024 USA Presidential election. Remember DO NOT commit this data to your repository!

**Step 1: Choose a small number of public datasets to work with (this is not your first assignment).**

Note that I will want you to include the information one would need to download such a dataset in the future so that a researcher might replicate your findings. Do not commit these data to your repository though.

### ****Step 2: Required Scripts****

Write **three separate Python scripts (or R or Julia or ...)**:

**Script 1 — Pure Python (No Pandas, No Polars)**

* Load the dataset using only the standard library.
* For each column and the dataset overall, compute:
  + Count
  + Mean (for numeric fields)
  + Minimum and maximum values
  + Standard deviation (optional if using only math module)
  + For non-numeric fields: unique value counts and most frequent values
* Perform the same analysis as above after aggregation by (“page\_id”) and again after aggregation by (“page\_id”, “ad\_id”) or other grouping columns in the even you are working with other datasets. For example, if your dataset includes a column for “state” that might be a candidate for grouping by.

**Script 2 — With Pandas**

* Use pandas to load and analyze the dataset.
* Replicate the same descriptive statistics as above using tools like:
  + DataFrame.describe()
  + value\_counts()
  + nunique()

**Script 3 — With Polars**

* Use pandas to load and analyze the dataset.
* Replicate the same descriptive statistics as above using tools like:
  + DataFrame.describe()
  + value\_counts()
  + nunique()

### ****Bonus Challenges – Create Supporting Visualizations of this data****

**Assume you have been tasked by a senior executive or researcher to produce a descriptive analysis of this dataset including a narrative of your finding that they might use as part of a presentation, or that you might directly present to these principals. Does any aspect of this dataset present a compelling story or narrative? If so, please elaborate on your findings**

* Create relevant visualizations for numeric and categorical columns.
  + Examples: histograms, bar charts, boxplots
  + Tools: matplotlib, seaborn, plotly (Python) or ggplot2, plotly ®

### ****Submission Instructions****

* Create a **public GitHub repository** titled: Task\_04\_Descriptive\_Stats
* Include the following:
  + pure\_python\_stats.py (potentially specific to each dataset)
  + pandas\_stats.py (potentially specific to each dataset)
  + polars\_stats.py (potentially specific to each dataset)
  + Any bonus scripts, notebooks or presentations
  + A brief README.md with:
    - Instructions to run your code
    - A short summary of findings or interesting insights
* **Do not include the dataset file** in your GitHub repo.
* Submit your repository link to [jrstrome@syr.edu](mailto:jrstrome@syr.edu) by: **2025-07-01**

**Time Reporting Requirement:**

**It is critical that you report your research progress via the Qualtrics survey. This is the main way we are tracking OPT activity for when we must report to the government. Please ensure you complete these check-ins in a timely manner.**

**https://syracuseuniversity.qualtrics.com/jfe/form/SV\_cDgnzM695AMx8d8**