Progress Report 1

Big Data Tools for Economists, Professor Khazra

For each section below, please write no more than 3 sentences (except tables). The entire report should not exceed one page to maximum one and a half. This report is designed to help you identify data issues **early** and to scaffold your progress toward the final project. A perfect score on this report does **not** imply that your analysis is correct or that it can be directly reused in the final project.

We will primarily grade these reports based on **effort, clarity, and completion**. You are strongly encouraged to discuss problems, uncertainties, and partial progress with the TAs. If you are unsure about any step, please bring your questions to the TAs during **Collaboration Hours**.

- IMPORTANT: In your submission files, please keep the **section** and **item** titles in bold *exactly as provided*. This will help us quickly locate the information and ensure consistent grading.
- If the section number and title does not match what we have here or is missing, you will lose the marks for that section.
- Write each section as short as possible, no more than 3 lines per section (excluding any tables reported). **Longer submissions will be penalized by up to 100%**.

Topic: Merge and GroupBy in Python

The goal of this report is to ensure that you can correctly **merge multiple datasets** and use **groupby operations** to generate summary statistics. Most empirical issues in applied work arise at the data construction stage. This report is meant to surface those issues early, well before the project.

You should already have selected a dataset (or datasets) from the provided list. This week's work should help you get started with your data and understand the variable you have or don't have.

1. Datasets and Motivation

Data Description

Describe the datasets you are using and why you need each of them:

- **Dataset List:** Briefly list all datasets you are using (names, sources).
- **Purpose of Each Dataset:** Explain what information each dataset provides and why it is necessary for your project.
- **Unit of Observation:** State the unit of observation for each dataset (e.g., individual-year, firm-quarter, county-year).

2. Merge Strategy

Merge Details

Explain how you merged (or plan to merge) the datasets:

- **Merge Keys:** Specify the variable(s) used for merging and justify why they are appropriate.
- **Type of Merge:** Indicate whether you used an inner, left, right, or outer merge.
- **Data Issues:** Mention any problems encountered (e.g., duplicates, missing keys, inconsistent identifiers).

3. Sample Size and Merge Diagnostics

Observations Before and After Merge

Carefully document how the merge affected your sample:

- **Pre-Merge Sizes:** Report the number of observations in each dataset before merging.
- **Post-Merge Size:** Report the number of observations in the final merged dataset.
- **Dropped Observations:** State whether any observations were dropped and explain why.
- **Selection Concerns:** Discuss whether dropped observations appear random or systematic.

Important Reminder: At no point should your dataset used for regressions, machine learning, or causal analysis fall below **a few thousand observations**. While aggregation is needed for exploratory analysis or visualization, your main analytical dataset must remain sufficiently large. Please talk to me if this is not clear or if your dataset does not allow for this.

4. GroupBy and Summary Statistics

Descriptive Analysis

Use `groupby` operations to summarize your data:

- **Grouping Variables:** Specify the variable(s) you grouped by (e.g., treatment status, time, region).
- **Summary statistics tables:** Report key variables in a summary statistics table (means, counts, or other relevant measures). This is better done using your final and merged dataset.
- **Interpretation:** Briefly explain what these summaries reveal about your data (1-2 lines, no more).

5. Reflection and Next Steps

Project Scaffolding

Reflect on what you learned from this week:

- **Data Viability:** Do your datasets and merges seem suitable for the final project?
- **Remaining Issues:** Identify unresolved data problems or concerns.
- **Planned Improvements:** Describe what you plan to fix or refine before moving forward.

Final Note: This report is a diagnostic and learning tool. Scoring 100% does **not** mean your analysis is correct, complete, or final-project ready. Use these weekly reports to iterate, identify weaknesses, and actively seek feedback from TAs during **Collaboration Hours**.