SIADS 593: Milestone I

Team Project Proposal

version 2022.07.27.1.CT

## **Proposal Title:** Identifying Alternative Proxies for Nowcasting US GDP

## 1. Team members

Please list your team members (2-3 max).

* Jan Nagtegaal (JJNAG)
* Aditya Sharma (ADITYSH)
* Evan Troutt (CTEVAN)

## 2. Project summary

Summarize your proposed project in a few sentences.

#### What is your proposed project and why are you proposing it?

#### What are the question(s) you want to answer, or goal you want to achieve?

| Introduction: The Problem (why)  * **Context:** Traditional quarterly GDP reports often lag, failing to capture rapid economic changes. * **Impact:** This leads to decision-making based on outdated data, affecting policymakers, investors, and economic analysts, and potentially impacting financial markets and policy development. * **Solution:** This project proposes to bridge this information gap by identifying and utilizing high-frequency data proxies for GDP components, providing more immediate insights into the country's economic health.  Project Scope and Objectives **Primary Goal:** To identify alternative proxies that can nowcast US GDP growth, offering a more current view of the economic status.  **Challenges Addressed:** Bridging the gap in current economic representation due to delayed GDP reports thus meeting the need for immediate economic indicators.  **Approach:** Methodical data analysis and proxy validation will be employed to provide a more accurate real-time economic assessment. **Key Questions:**  * Which high-frequency data sources can serve as effective proxies for GDP components such as Consumption, Investment, Government Expenditure, and Net Exports? * How can these data proxies be validated and integrated to accurately reflect current economic activities and correlate with historical GDP data? * What methodologies can be employed to ensure timely, relevant, and effective data-driven insights for nowcasting GDP growth?  Methodology  * **Exploratory Data Analysis (EDA):** Conduct comparative analysis to understand data characteristics and relationships among economic indicators. * **GDP and Economic Indicators Correlation:**   + Assess the correlation between GDP and various economic indicators using scatter plots, Pearson correlation coefficients, and heatmaps.   + Identify indicators most strongly correlated with GDP changes. * **Interaction Analysis and Time Series Analysis:** Explore how different indicators interact over time and integrate this with external data sources. * **Statistical Analysis:** Provide a comprehensive statistical summary and correlation analysis to evaluate the predictive power of different indicators. * **Economic Significance Evaluation:** Assess the impact of quarterly variables on GDP and their predictive power. * **Indicator Selection:** Discuss the relevance and coverage of indicators, ensuring diverse economic representation.  Resources and Constraints  * **Data Resources:** Availability of high-frequency data and historical GDP figures. * **Limitations:** Project timelines may affect the scope of analysis.  Assumptions and Risks  * **Data Quality and Relevance:** The assumption that available data is accurate, relevant, and timely. * **Model Predictive Power:** The risk that the developed model may not fully capture the complexities of the economy or may be subject to biases. |
| --- |

## 3. Datasets

#### Describe one primary dataset and at least one secondary dataset. If other secondary datasets will be used please describe them as well.

#### The proposed datasets should exhibit different features/columns and/or different access methods, e.g., \*.csv file, \*.json file, API retrieval, web scraping, etc. Different time periods, for example, with the same features/columns is not considered a different dataset. Remember, the focus of the project in this Milestone course is to give you the opportunity to practice your data manipulation skills, so feel free to challenge yourself.

#### If you're unsure if your data sets are "different enough" describe the datasets and request a review via the *#siads593\_[semester]\_001\_project* Slack channel.

#### **Please note:** all proposed datasets ***MUST*** be publicly available to all members of the class (students, instructors, course support personnel, etc.). Use of proprietary datasets for this project is ***not*** permitted.

## 3.1 Primary dataset description

Describe your primary dataset. How is the data collected and how will you access it? Please share what features in the dataset are relevant to your topic. At a minimum, include the following information:

#### Short description (i.e., 1-3 sentences) of its key features

#### Estimated size (in records and/or bytes)

#### Location (give the URL or other access method)

#### Format (CSV, JSON, etc.)

#### Access method (download, web scraping, API, etc.)

| Primary Dataset Description **Short Description:** The primary dataset is the "Table 1.1.5. Gross Domestic Product" from the U.S. Bureau of Economic Analysis. It comprises seasonally adjusted quarterly rates of the U.S. Gross Domestic Product (GDP) in billions of dollars. The data reflects the economic output of the United States and is crucial for analyzing economic trends and growth patterns.  **Key Features:**   * Seasonal adjustment of GDP values. * Presentation of data in quarterly rates by GDP component * GDP figures represented in billions of dollars.   **Estimated Size:** Approximately 0.51MB.  **Location:** Available at [U.S. Bureau of Economic Analysis](https://apps.bea.gov/iTable/?reqid=19&step=2&isuri=1&categories=survey&_gl=1*j1lvlb*_ga*MTk0MDMyMjk0MC4xNzA1NDk1NTk4*_ga_J4698JNNFT*MTcwNTQ5NTU5OC4xLjEuMTcwNTQ5NzA2MC42MC4wLjA.#eyJhcHBpZCI6MTksInN0ZXBzIjpbMSwyLDMsM10sImRhdGEiOltbImNhdGVnb3JpZXMiLCJTdXJ2ZXkiXSxbIk5JUEFfVGFibGVfTGlzdCIsIjUiXSxbIkZpcnN0X1llYXIiLCIxOTQ3Il0sWyJMYXN0X1llYXIiLCIyMDIzIl0sWyJTY2FsZSIsIi05Il0sWyJTZXJpZXMiLCJRIl1dfQ==). ([BEA](https://apps.bea.gov/iTable/?reqid=19&step=2&isuri=1&categories=survey&_gl=1*j1lvlb*_ga*MTk0MDMyMjk0MC4xNzA1NDk1NTk4*_ga_J4698JNNFT*MTcwNTQ5NTU5OC4xLjEuMTcwNTQ5NzA2MC42MC4wLjA.#eyJhcHBpZCI6MTksInN0ZXBzIjpbMSwyLDMsM10sImRhdGEiOltbImNhdGVnb3JpZXMiLCJTdXJ2ZXkiXSxbIk5JUEFfVGFibGVfTGlzdCIsIjUiXSxbIkZpcnN0X1llYXIiLCIxOTQ3Il0sWyJMYXN0X1llYXIiLCIyMDIzIl0sWyJTY2FsZSIsIi05Il0sWyJTZXJpZXMiLCJRIl1dfQ==))  **Format:** CSV  **Access Method:** The dataset can be accessed and downloaded directly from the website.  **Relevance to Topic:** The dataset's detailed information on U.S. GDP over several years is integral to the project's goal of nowcasting GDP growth. The data's granularity and time-series nature will allow for comprehensive analysis and identification of trends, making it pivotal for the project's success. |
| --- |

## 3.2 Secondary dataset(s) description

Describe your secondary dataset(s). How is the data collected and how will you access it? Please share what features in the dataset(s) are relevant to your topic and describe the data types you’re expecting. At a minimum, for each secondary dataset include the following information:

#### Short description (i.e., 1-3 sentences) of its key features

#### Estimated size (in records and/or bytes)

#### Location (give the URL or other access method)

#### Format (CSV, JSON, etc.)

#### Access method (download, web scraping, API, etc.)

| Secondary DatasetDataset 1: Federal Reserve Economic Data (FRED) - Quarterly Data **Short Description:** This dataset is from the Federal Reserve Bank of St. Louis's FRED macroeconomic database. It includes a wide range of economic data points available at a quarterly frequency, with a focus on US GDP data. This dataset is essential for providing contextual economic indicators that complement the primary GDP data.  **Key Features:**   * Comprehensive collection of quarterly macroeconomic data. * Focus on US GDP data among various economic indicators.   **Estimated Size:** Approximately 0.4MB.  **Location:** Available at [Federal Reserve Economic Data - Quarterly](https://research.stlouisfed.org/econ/mccracken/fred-databases/). <https://research.stlouisfed.org/econ/mccracken/fred-databases/>  **Format:** CSV  **Access Method:** The dataset can be accessed and downloaded directly from the FRED website.  **Relevance to Topic:** The quarterly frequency of this dataset aligns with the primary dataset's timeframe. It offers additional economic indicators that can be analyzed alongside the primary GDP data to enhance the understanding of economic trends and contribute to the nowcasting of GDP growth. Dataset 2: Federal Reserve Economic Data (FRED) - Monthly Data **Short Description:** This dataset, also from the FRED macroeconomic database, includes a variety of economic data points available at a monthly frequency. It is assumed that some or all of these data points can serve as proxies for predicting GDP growth, providing a more granular view of economic trends.  **Key Features:**   * Diverse range of monthly economic data points. * Potential proxies for GDP growth prediction.   **Estimated Size:** Approximately 0.6MB.  **Location:** Available at [Federal Reserve Economic Data - Monthly](https://research.stlouisfed.org/econ/mccracken/fred-databases/). <https://research.stlouisfed.org/econ/mccracken/fred-databases/>  **Format:** CSV  **Access Method:** The dataset can be accessed and downloaded directly from the FRED website.  **Relevance to Topic:** The **monthly** frequency of this dataset offers a finer temporal resolution compared to the primary dataset, potentially revealing more immediate economic trends. This granularity will be instrumental in identifying more immediate proxies for GDP nowcasting. |
| --- |

## 3.3 [Yes] Affirm: datasets are public.

Please write YES in the above box to confirm that your primary and secondary datasets are accessible and available to your classmates and the instructional team.

## 4. Cleaning and manipulation

Describe how you will need to manipulate your datasets: how will you handle missing or anomalous data? How will you join your primary and secondary datasets? What cleaning and manipulation challenges, if any, do you anticipate?

| Data Cleaning and Consistency Checks  * **Initial Cleaning:** Handle missing values and outliers. Outliers will be identified using the Interquartile Range (IQR) method and additional methods like visual analysis and standard deviation. * **Data Standardization:** Standardize date formats using pandas datetime functions and ensure consistent data types across datasets. Convert quarterly data columns from object to float for numerical analysis. * **Handling Anomalies:** Apply Z-score and temporal analysis to robustly detect anomalies.  Data Transformation  * **Frequency Alignment:** The significant challenge is the conversion of different data frequencies. Aggregating monthly data to quarterly requires careful consideration to avoid data distortion. * **Stationarity Checks:** Address non-stationary factors using logarithmic transformations or differencing. * **Seasonal Adjustments:** Employ smoothing techniques to reduce seasonal variations. * **Disaggregation:** We’ll need to disaggregate lower frequency GDP data which is available on a quarterly frequency into higher frequency data of economic indicators which are available on a monthly basis * **Mapping to FRED Definitions:** Align secondary dataset categories with FRED's definitions for consistency. * **Format Harmonization:** Ensure the data formats of the BEA and FRED datasets are compatible. * **Date Standardization:** Modify the monthly datasets to have dates compatible with the other datasets. * **Setting Hierarchies:** Organize the BEA dataset to reflect the hierarchy in GDP calculation.  Data Integration  * **Dataset Merging:** merge the datasets into a single pandas dataframe. * **Data Consistency Checks:** It involves not just aligning data points but also ensuring that the combined dataset accurately represents economic realities without duplication or loss of crucial information.  Anticipated Challenges  * **Data Quality and Reliability:** Critical to ensure the accuracy of data during the cleaning and standardization processes. * **Complex Frequency Conversion:** Monthly to quarterly data conversion involves intricate handling to maintain data integrity. |
| --- |

## 5. Analysis

Describe any analyses you plan to undertake. For each, please give the technique or approach and briefly explain what you expect to learn from it.

| 1. Statistical Analysis  * **Correlation Analysis**   + **Technique:** Employ Pearson correlation coefficients, scatter plots, and heatmaps.   + **Objective:** Identify the economic indicators most strongly correlated with GDP to reveal key factors influencing GDP movements.   + **Possible Visualization:** Heatmaps and scatter plots to visually represent correlation strengths and discern patterns. * **Time-Series Decomposition**   + **Technique:** Decompose the GDP time series into trend, seasonality, and residuals.   + **Objective:** Discover underlying patterns in GDP data, crucial for understanding the economic cycle.   + **Possible Visualization:** Time-series plots to clearly illustrate separated components**.** * **Principal Component Analysis (PCA)**   + **Technique:** Use PCA to reduce dataset dimensionality.   + **Objective:** Identify and concentrate on the most significant indicators, eliminating redundancy and focusing analysis on impactful variables.   + **Possible Visualization:** Biplots or Scree plots to interpret the variance captured by each principal component.  2. Comparative Analysis  * **Lead and Lag Analysis**   + **Technique:** Investigate temporal relationships between GDP and its indicators.   + **Objective:** Determine potential leading or lagging indicators for GDP, assisting in understanding economic cycles.   + **Possible Visualization:** Lag or cross-correlation plots to determine leading/lagging characteristics. * **Stationarity Testing**   + **Technique:** Apply the Augmented Dickey-Fuller test.   + **Objective:** Ensure time series data is suitable for use in forecasting models by confirming stationarity.   + **Possible Visualization:** Time-series plots pre- and post-transformation. * **Economic Indicators Correlation**   + **Technique:** Conduct visual analysis using scatter plots and heatmaps.   + **Objective:** Understand the strength and nature of relationships between GDP and various economic indicators.   + **Possible Visualization:** Scatter Plot Matrix (SPLOM) for exploring inter-variable relationships. * **Impact of Specific Events**   + **Technique:** Analyze GDP and indicator trends around significant economic events using annotated time-series plots.   + **Objective:** Quantify the effect of major events like financial crises on GDP and related indicators.   + **Possible Visualization:** Annotations on time-series plots  3. Evaluation  * **Variable Selection and Reduction:**   + **Technique:** Use outcomes from PCA and correlation analysis to select a subset of indicators that are most relevant and non-redundant.   + **Objective:** Reduce overfitting and complexity in the model by focusing on a concise set of variables that accurately represent economic dynamics. * **Model Evaluation and Uncertainty Assessment:**   + **Technique:** Apply methods like bootstrapping or Monte Carlo simulations.   + **Objective:** Assess the model's accuracy and the uncertainty associated with GDP predictions, ensuring robustness and reliability.   + **Possible Visualization:** Overlay plots of predicted vs. actual GDP values to visually assess model performance. * **Regression Analysis (without Modeling):**   + **Technique:** Conduct linear regression on selected significant variables as an evaluative measure.   + **Objective:** Understand the influence of each economic indicator on GDP, using regression coefficients as indicative weights.   + **Possible Visualization:** Regression line plots on scatter plots to show relationships between GDP and indicators. |
| --- |

## 6. Visualizations

Describe in 1-3 sentences at least **two** data visualizations that you plan to create. Include the chart type (e.g. bar chart, scatterplot, SPLOM, etc.) as well as the variables (features) you intend to plot.

| * **Scatter Plots with Trend Lines for Economic Indicators and GDP:** This visualization will consist of individual scatter plots for each economic indicator, such as inflation rate, consumer spending, and unemployment rate, plotted against GDP. Each plot will include a trend line to visually represent the relationship and correlation strength between the indicator and GDP. The goal is to visually gauge the extent to which each economic variable correlates with GDP movements, offering insights into their relative influence on the economy. * **Line Graph Comparing Actual vs. Predicted GDP with Confidence Intervals:** This graph will plot the actual GDP data alongside the predicted GDP values over time, presented as a line graph. To illustrate the uncertainty in the predictions, confidence intervals or prediction bands will be included around the predicted values. This visualization aims to not only compare the predicted GDP against actual figures but also to visually communicate the range of uncertainty associated with the predictions, providing a clearer understanding of the model's reliability. * **Scatter Plot Matrix (SPLOM) for Economic Indicators and GDP:** This visualization will encompass multiple scatter plots in a matrix format to display relationships among various economic indicators and GDP. Key variables like GDP growth, consumption, investment, government expenditure, and net exports will be plotted. This comprehensive view will allow us to visually assess correlations and patterns across multiple indicators simultaneously. * **Time-Series Decomposition Plot:** A time-series plot will be used to decompose the GDP data into its trend, seasonality, and residual components. This will involve plotting the original GDP series alongside its decomposed elements. The objective is to visually dissect and understand the underlying patterns within the GDP data, such as long-term trends and cyclical behaviors. * **Heatmap for Correlation Analysis:** A heatmap will be created to visualize the Pearson correlation coefficients between various economic indicators and GDP. This chart will plot indicators such as consumer sentiment, inflation rate, unemployment rate, and others against GDP, displaying the strength of their correlations through color intensity. This visualization will effectively highlight which indicators have the most significant relationships with GDP. |
| --- |

## 

## 7. Ethical considerations

Does your choice of data raise any ethical issues? If so, briefly describe the concern and how you plan to mitigate it.

| While our project primarily utilizes publicly available data, thereby minimizing direct privacy concerns, there are still critical ethical considerations to address:  Risk of Misinterpretation:   * Given the complexity and sensitivity of economic data, there is a risk of misinterpreting the information or drawing inaccurate conclusions. This can be particularly consequential if our findings are used to inform economic policies or investment decisions. * **Mitigation Strategy:** We will ensure rigorous statistical analysis and employ validation techniques to enhance the accuracy of our interpretations. Moreover, we will provide clear explanations of our methodology and the assumptions underlying our analysis to prevent misinterpretation.   Causation vs. Correlation:   * A significant ethical concern is the potential to imply causation where there is only evidence of correlation. Economic indicators and GDP are influenced by a myriad of factors, and suggesting direct causality without robust evidence can be misleading. * **Mitigation Strategy:** In our reporting and communication of results, we will explicitly clarify that our findings indicate correlation and not causation. We will stress that while certain indicators may be strongly correlated with GDP, this does not necessarily imply a cause-and-effect relationship.   Communication of Limitations:   * It's crucial to acknowledge and communicate the limitations of our model and its predictions. Overstating the model's predictive power or failing to disclose its limitations can lead to overreliance on its outputs. * **Mitigation Strategy:** We will transparently communicate the scope, limitations, and uncertainties of our model. This includes discussing the potential margin of error and the conditions under which our model's predictions are most reliable. |
| --- |

## 8. Contributions

Indicate the contribution that each team member will make to the project.

| Team Members:  Person name 1 (JJNAG)  Person name 2 (ADITYSH)  Person name 3 (CTEVAN) **Draft Project Plan**Phase 1: Initiation and Planning  * **All Members:** Formulation of the project problem, its context, and impact. Develop the initial project outline and objectives. * **Mentor:** Assist in defining the project scope and objectives. Outline key questions and anticipated challenges. * **All Members:** Refine Project from on Mentor feedback.  Phase 2: Data Acquisition and Cleaning  * **JJ:** Responsible for initial data cleaning of both primary and secondary datasets, including handling missing values and outliers. * **Aditya:** Lead the data integration process, ensuring consistency and accuracy in the combined dataset. * **Evan:** Assist in data cleaning, particularly in standardizing data formats and types.  Phase 3: Exploratory Data Analysis (EDA)  * **Evan & JJ:** Conduct statistical analysis, including correlation analysis and PCA. Develop initial visualizations like heatmaps and scatter plots. * **Aditya:** Perform time-series decomposition and stationarity testing. Visualize the findings using time-series plots.  Phase 4: In-Depth Comparative Analysis  * **JJ:** Undertake lead and lag analysis to explore temporal relationships. Create corresponding visualizations. * **Evan:** Analyze economic indicators' correlation with GDP and assess the impact of significant events.  Phase 5: Data Synthesis and Evaluation  * **Aditya:** Evaluate the model's accuracy and uncertainty. Apply bootstrapping or Monte Carlo simulations as necessary. * **JJ:** Use outcomes from PCA and correlation analysis for variable selection and reduction.  Phase 6: Regression Analysis and Insights  * **Evan:** Conduct regression analysis to understand the influence of economic indicators on GDP. Visualize using regression plots. * **Aditya:** Compare regression analysis findings with other analyses for consistency and accuracy.  Phase 7: Documentation and Reporting  * **JJ:** Lead the compilation of findings, analyses, and visualizations into the final project report. * **Evan and Aditya:** Contribute to writing and reviewing the report, ensuring clarity in methodology, results, and limitations.  Regular Meetings and Collaboration Participate in weekly meetings to discuss progress, address challenges, and review the effectiveness of analyses and visualizations. Regularly review each other’s contributions for quality and coherence. Collaboration and Version Control  * **GitLab:** Our primary platform for project tracking and collaboration. GitLab will be used to manage tasks, track progress, and facilitate team communication. * **GitHub:** We will utilize a GitHub repository for storing all project documentation, datasets, and to maintain version control. This ensures that our project's history is well documented and easily accessible for all team members.  Development and Analysis Environment  * **Google Colab:** For interactive development and collaborative analysis, we will use Google Colab. Its cloud-based environment is ideal for sharing Jupyter notebooks and working together in real-time. * **Virtual Environments:** To maintain consistency across our development work, we will set up virtual environments. This approach guarantees that all notebooks, data imports, libraries, and dependencies are reliable, consistent, and reproducible, regardless of the individual team member’s local setup. Google Colab does not directly support virtual environments since it runs in a cloud-based environment. We will ensure consistency by installing the required packages at the beginning of your Colab notebooks using the requirements.txt file from your GitHub repository to install the same dependencies in Colab.  Data Analysis and Visualization  * **Python:** The primary programming language for our data analysis. |
| --- |