SIADS 593: Milestone I

Team Project Proposal

version 2022.07.27.1.CT

## **Proposal Title:** Nowcasting Consumer Expenditure: Uncovering Reliable Proxies for Consumer Spending Behaviour.

## 1. Team members

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

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

## 2. Project summary

Summarise 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 * **Impact:** This leads to decision-making based on outdated data, affecting policymakers, investors, and economic analysts and potentially impacting financial markets and policy development.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 utilising high-frequency data proxies for GDP components, providing more immediate insights into the country's economic health.  Introduction: The Problem (Why)  * **Context**: Traditional quarterly GDP reports, especially the consumer expenditure component, often need to catch up, failing to capture the nuances of rapid economic shifts. * **Impact**: This delay leads to decisions based on outdated consumer spending data, affecting many stakeholders including policymakers, retailers, financial analysts, and investors. * **Solution**: The project aims to bridge this information gap by identifying and utilising high-frequency data proxies specifically for consumer expenditure, providing more immediate insights into consumer behaviour and spending trends.  Project Scope and Objectives (What)  * **Primary Goal**: To identify and validate alternative proxies that can nowcast consumer expenditure in the U.S., offering a real-time snapshot of consumer spending behaviours. * **Challenges**: Addressing the lag in consumer spending data reporting and filling the need for immediate and relevant economic indicators. * **Approach**: A systematic approach involving data analysis, proxy validation, and integration to provide an accurate real-time assessment of consumer expenditure. * **Key Questions**:   + What high-frequency data sources can act as reliable proxies for consumer spending behaviour?   + How can these data proxies be validated and correlated with traditional measures of consumer expenditure?   + What methodologies can ensure the proxies provide timely and relevant insights into consumer spending trends?  Methodology  * **Exploratory Data Analysis (EDA)**: Conduct a comparative analysis to understand the characteristics of data related to consumer spending and its relationship with other economic indicators. * **Consumer Expenditure and Economic Indicators Correlation**:   + Use scatter plots, Pearson correlation coefficients, and heat maps to assess the correlation between traditional consumer expenditure data and potential proxies.   + Identify indicators most strongly correlated with changes in consumer spending. * **Interaction Analysis and Time Series Analysis**: Explore how consumer spending proxies interact over time and how they can be integrated with other economic data sources. * **Statistical Analysis**: Provide a comprehensive statistical summary and correlation analysis to evaluate the predictive power of different proxies for consumer expenditure. * **Indicator Selection**: Focus on the selection of relevant and diverse proxies for consumer expenditure, ensuring they represent a comprehensive view of consumer behavior.  Resources and Constraints  * **Data Resources**: Utilization of high-frequency data such as social media trends, online retail data, and consumer confidence indices, alongside traditional consumer expenditure data. * **Limitations**: Constraints include project timelines and the availability of timely high-frequency data.  Assumptions and Risks  * **Data Quality and Relevance**: Assumption that the selected high-frequency data sources are accurate, relevant, and timely. * **Model Predictive Power**: The risk that the proxies may not fully capture the complexities of consumer behavior or may be subject to biases in representing consumer expenditure. |
| --- |

## 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 "Table 1.1.5. Gross Domestic Product" from the U.S. Bureau of Economic Analysis. It comprises seasonally adjusted quarterly U.S. Gross Domestic Product (GDP) rates in billions of dollars. The data reflects the economic output of the United States and is crucial for analysing economic trends and growth patterns.  **Key Features:**   * Seasonal adjustment of GDP values. * Presentation of data in quarterly rates by GDP component * GDP figures are 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 is readily available and can be easily accessed and downloaded directly from the U.S. Bureau of Economic Analysis 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 consumption. 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)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 and monthly frequency, with a focus on US GDP data. This dataset Contains a wide range of economic data points including consumer spending indicators, a crucial component of the Gross Domestic Product (GDP). **Relevance to Topic:** The **monthly** frequency of this dataset offers a finer temporal resolution than the primary dataset, potentially revealing more immediate economic trends. This granularity will be instrumental in identifying more immediate proxies for nowcasting.   * **Key Features**: Diverse range of economic data points, both at a monthly and quarterly frequency. * **Estimated Size**: 0.4MB (Quarterly), 0.6MB (Monthly). * **Location**: Available at [FRED - Quarterly](https://research.stlouisfed.org/econ/mccracken/fred-databases/) and [FRED - Monthly](https://research.stlouisfed.org/econ/mccracken/fred-databases/). <https://research.stlouisfed.org/econ/mccracken/fred-databases/> * **Format**: CSV. * **Access Method**: Direct download. * **Relevance**: Complements the primary dataset with additional economic indicators, useful for cross-referencing and correlation analysis.   **E-commerce Sales Data**   * **Short Description**: Aggregate data on online retail sales from major e-commerce platforms. * **Key Features**: Sales figures, product category trends. * **Estimated Size**: Varies based on selected sources. * **Location**: Available from public datasets or through APIs of platforms like [Amazon](https://developer.amazon.com/). * **Format**: CSV, JSON. * **Access Method**: API retrieval or web scraping. * **Relevance**: Direct indicator of consumer spending, especially relevant in the digital age. |
| --- |

## 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?

| Initial Cleaning and Standardization  * **Handling Missing Values and Outliers**: Use methods like the Interquartile Range (IQR) and Z-score analysis to identify and address outliers and missing values in consumer spending datasets. * **Standardisation of Data Types and Formats**: Utilize pandas for standardising data formats, ensuring consistent data types across datasets and, for example, converting data columns to appropriate numerical formats for analysis.  Data Transformation and Alignment  * **Frequency Alignment**: Harmonize different data frequencies, especially aligning higher-frequency data (e.g., monthly Twitter sentiment data) with the lower-frequency primary economic data (quarterly consumer spending figures). * **Format Harmonization and Integration**: Ensure compatibility in data formats across traditional and non-traditional economic datasets (like JSON format from Twitter API and CSV from economic databases). * **Temporal Analysis**: Implement time-series analysis techniques to understand consumer spending trends over time and correlate them with high-frequency proxy data.  Handling Anomalies and Data Quality  * **Anomaly Detection**: Apply robust statistical methods to detect anomalies in consumer spending patterns, considering both traditional and non-traditional datasets. * **Data Quality Assurance**: Perform consistency checks and validate the accuracy and relevance of the primary economic data and the high-frequency proxy data.  Data Integration and Aggregation  * **Merging Datasets**: Combine datasets into a single framework using pandas, ensuring seamless integration of different data sources. * **Data Aggregation**: Aggregate high-frequency data (like daily or monthly data points) to align with the quarterly consumer expenditure data for comparative analysis. * **Disaggregation**: In some cases, disaggregate quarterly data to match with higher-frequency indicators, applying appropriate statistical techniques to maintain data integrity.  Seasonal Adjustments and Stationarity Checks  * **Seasonal Adjustments**: Apply smoothing techniques to reduce seasonal variations, especially in consumer spending data, which can be highly seasonal. * **Stationarity Checks**: Use methods like the Augmented Dickey-Fuller test to ensure the stationarity of time-series data, which is crucial for reliable correlation analysis.  Addressing Non-Stationary Factors and Data Complexity  * **Non-Stationary Factors**: Address non-stationary elements in economic indicators and high-frequency data proxies, potentially using techniques like differencing or transformation. * **Complexity in Frequency Conversion**: Tackle challenges in converting and aligning data of varying frequencies, ensuring data integrity and relevance preservation.  Anticipated Challenges  * **Data Quality and Reliability**: Ensuring the reliability and accuracy of data during cleaning and standardisation, particularly with diverse data sources. * **Complex Data Integration**: Integrating traditional economic data with non-traditional, high-frequency data presents unique challenges in maintaining consistency and accuracy. * **Real-Time Data Handling**: Managing and processing real-time data streams from sources like social media and Google Trends, ensuring timely and relevant analysis. |
| --- |

## 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 heat maps to measure the strength and direction of relationships.**Objective**: To identify which high-frequency data sources, such as e-commerce data, social media sentiment, and search trends, show a strong correlation with traditional consumer expenditure data.**Relevance to Scope**: This analysis directly addresses the need to identify reliable proxies for consumer spending behaviour, which is critical to effective nowcasting.**Possible Visualization**: Use heatmaps to visually represent correlation strength across multiple variables and scatter plots to depict individual relationships.2. Consumer Confidence Index Analysis**Technique**: Apply correlation and comparative analysis to examine the link between consumer sentiment and spending.**Objective**: Determine how changes in consumer confidence indices (CCI) correlate with actual consumer spending data fluctuations.**Approach**: Cross-reference CCI data with actual spending trends and use statistical methods to assess the relationship. This analysis will help determine if CCI can be a predictive indicator for spending behaviour.**Possible Visualization**: Create scatter plots or line graphs to depict the trend correlation between CCI and consumer spending over time.3. Time-Series Decomposition**Technique**: Break down consumer spending time series data into its constituent components: trend, seasonality, and residuals.**Objective**: Identify underlying patterns and cyclicality in consumer spending, which are crucial for understanding behaviour shifts.**Relevance to Scope**: This decomposition aids in dissecting the cyclical and seasonal aspects of consumer spending, aligning with the goal of nowcasting.**Possible Visualization**: Generate time-series plots that clearly illustrate each decomposed component.4. Comparative and Temporal Analysis**Lead and Lag Analysis**:**Technique**: Investigate the temporal relationships between traditional consumer spending data and high-frequency proxies to identify leading or lagging trends.**Objective**: Pinpoint indicators that could potentially forecast or trail behind consumer spending patterns.**Possible Visualization**: Use lag or cross-correlation plots to identify temporal alignments or disparities.**Stationarity Testing**:**Technique**: Implement the Augmented Dickey-Fuller test to assess the stationarity of time-series data.**Objective**: Ensure the reliability of time-series data, a crucial step before any further time-series analysis.**Possible Visualization**: Display time-series plots before and after any data transformations to illustrate the achievement of stationarity.**Consumer Behavior Indicators Correlation**:**Technique**: Conduct a detailed visual analysis using scatter plots and heatmaps to examine relationships.**Objective**: Explore how consumer spending correlates with various high-frequency proxies, uncovering strength and nature of these relationships.**Possible Visualization**: Develop a Scatter Plot Matrix (SPLOM) for an in-depth inter-variable relationship exploration.5. Proxy Evaluation and Variable Selection**Variable Selection and Reduction**:**Technique**: Analyze outcomes from the correlation analysis to narrow down the most relevant and impactful proxies for consumer spending.**Objective**: Focus on a select group of proxies that accurately reflect trends and shifts in consumer spending, reducing complexity and enhancing model accuracy.**Possible Visualization**: Overlay plots contrasting trends between chosen proxies and actual consumer expenditure data to validate their representativeness.**Model Evaluation and Uncertainty Assessment**:**Technique**: Implement techniques like bootstrapping or Monte Carlo simulations to evaluate the model's effectiveness.**Objective**: Assess the robustness and reliability of the proxy selection in accurately representing consumer expenditure trends.**Possible Visualization**: Utilize overlay plots to compare model predictions against actual data, highlighting the model's performance and uncertainty margins.6. Regression Analysis for Proxy Influence**Technique**: Perform linear regression analysis to determine the influence of each selected proxy on consumer spending.**Objective**: Quantitatively assess how each proxy impacts consumer spending, using regression coefficients to gauge their relative influence.**Possible Visualization**: Create regression line plots on scatter plots, showcasing the relationship between each proxy and consumer spending. |
| --- |

## 6. Visualisations

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.

| Heatmaps: To represent the strength of correlation between high-frequency data sources, e.g. e-commerce data and traditional consumer expenditure data.  Scatter Plots:For depicting individual relationships between variables such as consumer confidence indices (CCI) and actual consumer spending, as well as showcasing the relationship between each proxy (from the regression analysis) and consumer spending.  Line Graphs:To illustrate the trend correlation between consumer confidence indices and actual consumer spending over time.  Time-Series Plots:Used to demonstrate each component (trend, seasonality, residuals) from the time-series decomposition of consumer spending data. Also, it will display data before and after transformations in stationarity testing.  Lag or Cross-Correlation Plots: For identifying temporal alignments or disparities between traditional consumer spending data and high-frequency proxies during lead and lag analysis.  Scatter Plot Matrix (SPLOM):To explore in-depth inter-variable relationships in the consumer behavior indicators correlation analysis.  Overlay Plots:To contrast trends between chosen proxies and actual consumer expenditure data in the proxy evaluation and variable selection phase and compare model predictions against actual data in model evaluation and uncertainty assessment.  Regression Line Plots on Scatter Plots: To visualise the influence of each selected proxy on consumer spending as part of the regression analysis for proxy influence. |
| --- |

## 

## 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 utilises publicly available data, thereby minimising direct privacy concerns, there are still critical ethical considerations to address:  Privacy and Data Anonymity:   * **Concern**: Using social media and e-commerce data can raise privacy concerns, especially if individual user data is identifiable. * **Mitigation**: Ensure that all data used is anonymized and aggregated, with no possibility of tracing back to individual users. Adhere strictly to data use policies of the platforms and datasets.   Data Bias and Representation:   * **Concern**: Social media sentiment might not be representative of the entire population, leading to biased insights. Similarly, e-commerce data might over-represent certain demographics. * **Mitigation**: Acknowledge the limitations of the data in terms of representativeness. Supplement with more comprehensive datasets where possible. Use statistical methods to adjust for known biases.   Consent and Data Usage:   * **Concern**: The use of publicly available data, especially from social media, might not align with the original consent provided by users. * **Mitigation**: Use only publicly available data where users have consented to public sharing. Avoid scraping data in ways that violate the terms of service of the platforms.   Risk of Misinterpretation:   * Given the complexity and sensitivity of economic data, there is a risk of misinterpreting the information or drawing inaccurate conclusions. * **Mitigation Strategy:** We will ensure rigorous statistical analysis and employ validation techniques to enhance the accuracy of our interpretations. Moreover, we will clearly explain 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. Many factors influence economic indicators and consumer behaviour, and suggesting direct causality without robust evidence can be misleading. * **Mitigation Strategy:** In reporting and communicating results, we will explicitly clarify that our findings indicate correlation, not causation. We will stress that while certain indicators may be strongly correlated with consumer behaviour, 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 communicate our model’s scope, limitations, and uncertainties transparently. 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**: Refine the project's focus on consumer expenditure. Establish a clear context and impact for the study. Develop a detailed project outline with specific objectives centred around identifying proxies for consumer spending. * **Mentor**: Guide the team in narrowing the project scope. Help outline key questions and challenges specific to the project proposal. * **All Members**: Adjust the project based on mentor feedback, ensuring a concentrated approach on consumer spending.  Phase 2: Data Acquisition and Cleaning  * **JJ**: Focus on cleaning primary datasets related to consumer expenditure, addressing missing values and outliers. * **Aditya**: Lead the integration of secondary datasets, including e-commerce data, social media sentiment, and consumer confidence indices, ensuring consistency and relevance. * **Evan**: Assist in data cleaning, emphasising standardising formats and aligning different data frequencies, which is crucial for time-series analysis.  Phase 3: Exploratory Data Analysis (EDA)  * **Evan & JJ**: Conduct correlation analysis to identify relationships between consumer spending and high-frequency data sources. Develop initial visualisations, such as heat maps, to illustrate these correlations. * **Aditya**: Focus on time-series decomposition of consumer spending data to reveal underlying patterns. Visualise these components using time-series plots.  Phase 4: In-Depth Comparative Analysis  * **JJ**: Perform lead and lag analysis with consumer spending data and high-frequency proxies. Create visualisations like lag or cross-correlation plots to identify predictive trends. * **Evan**: Analyze the correlation between consumer spending and consumer confidence indices and its implications on spending patterns.  Phase 5: Data Synthesis and Evaluation  * **Aditya**: Evaluate the effectiveness of selected proxies using methods like bootstrapping. Assess the accuracy and uncertainty of these proxies in representing consumer spending trends. * **JJ**: Utilize outcomes from correlation analysis for variable selection, focusing on the most indicative proxies of consumer spending.  Phase 6: Regression Analysis and Insights  * **Evan**: Conduct regression analysis to determine the impact of each proxy on consumer spending. Visualise findings using regression line plots on scatter plots. * **Aditya**: Compare findings from regression analysis with correlation analysis for consistency and to validate the proxies.  Phase 7: Documentation and Reporting  * **JJ**: Compile the findings, analyses, and visualisations into a comprehensive final report. * **Evan and Aditya**: Contribute to writing and reviewing the report, ensuring clarity in methodology, results, and implications for consumer spending nowcasting.  Regular Meetings and Collaboration Participate in weekly meetings to discuss progress, address challenges, and review the effectiveness of analyses and visualisations. 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 utilise a GitHub repository to store all project documentation datasets and maintain version control. This ensures that our project's history is well-documented and easily accessible to 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:** We will set up virtual environments to maintain consistency across our development work. 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. |
| --- |