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.

#### 1.1. Introduction: The Problem (Why)

The current quarterly GDP reports lag in reflecting the dynamic changes in the economy, impacting decision-makers who rely on timely economic data. This project is devised to mitigate this issue by identifying high-frequency, readily updated data proxies that offer quicker insights into consumer expenditure patterns.

#### 1.2. Project Scope and Objectives (What)

The project's primary objective is to systematically identify, harmonise, and validate high-frequency data sources as proxies for real-time tracking of consumer expenditure in the United States. The goal is to refine these proxies to provide more immediate data on consumer spending habits, thus bridging the gap caused by the delayed reporting of official GDP figures.

#### Key Questions:

* Which high-frequency data sources can serve as accurate proxies for consumer spending?
* How can we validate these proxies against established measures of consumer expenditure?
* What techniques can we employ to ensure these proxies offer immediate and reliable insights into current consumer spending trends?
* How will we address potential discrepancies between different data sources in terms of scale, units, or reporting standards?
* Are there any unforeseen challenges in harmonizing data frequencies (monthly vs. quarterly) that could impact the accuracy of our analysis?
* How can we ensure the economic relevance of our findings, beyond statistical correlations?
* What contingency plans do we have for dealing with data anomalies or irregularities that might skew our analysis?

#### 1.3. Methodology

The methodology is designed to focus on data preparation and validation:

* **Exploratory Data Analysis (EDA)**: To understand the characteristics and quality of the high-frequency monthly indicators and their initial relationships to consumer spending.
* **Data Harmonization**: To transform and align the monthly indicators with the quarterly GDP data, using log transformations and adjustments for seasonality and rate of change.
* **Proxy Validation**: To establish a correlation with established measures of consumer spending through statistical analysis, ensuring that the proxies are reliable and relevant.

#### 1.4. Assumptions

**Data Quality and Relevance:** We operate under the assumption that the high-frequency data from FRED and other sources accurately reflect current economic trends and consumer sentiments. However, there is an inherent risk of data bias or inaccuracy, which could impact the reliability of our findings.

**Predictive Power and Relevance:** While we aim to identify effective proxies for consumer expenditure, there's a risk that these proxies may not fully capture the complexities of consumer behaviour or may not adapt swiftly to sudden economic shifts.

**External Factors:** The project also assumes a stable economic environment. Sudden external shocks (like global events or policy changes) could significantly affect consumer behaviour, potentially reducing the predictive accuracy of our proxies.

#### 2. 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.

**Relevance:** 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.

**Data frequency:** The data reflecting the economic output of the United States is crucial for analyzing economic trends and growth patterns. The presentation of data is done quarterly by the GDP component.

**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.

#### 3 Secondary Datasets

#### Federal Reserve Economic Data (FRED)

**Short Description:** This dataset is sourced from the Federal Reserve Bank of St. Louis's FRED macroeconomic database. It contains a variety of economic data points available at monthly intervals, with a particular focus on US GDP data. The data covers consumer spending indicators, a crucial component of the Gross Domestic Product (GDP).

**Relevance**: Complements the primary dataset with additional economic indicators, useful for cross-referencing and correlation analysis.

**Data frequency:** The monthly frequency of this dataset provides a more detailed temporal resolution than the primary dataset, which may reveal more immediate economic trends. This granularity will be useful in identifying more immediate proxies for nowcasting.

**Estimated Size**: 0.6MB

**Location**: <https://research.stlouisfed.org/econ/mccracken/fred-databases/>

**Format**: CSV.

**Access Method**: Direct download.

#### 4. Cleaning and Manipulation

#### 4.1 Initial Data Preparation

**Handling Missing Values**: Utilize median imputation for missing values, as it's less influenced by outliers and provides a more representative central tendency.

**Outliers and Anomalies**: Apply Interquartile Range (IQR) or Z-score analysis to identify and address outliers. This step ensures the integrity of data by minimizing the impact of extreme values.

**Data Type Standardization**: Use Python's Pandas library to standardise data formats and types across datasets. This step is crucial to ensure consistency, particularly when dealing with various formats like percentages, counts, and currencies.

#### 4.2 Data Harmonization and Transformation

**Log Transformation for Monthly Data**: Implement logarithmic transformations to stabilize the variance in monthly data that exhibit exponential growth or large fluctuations. This step is particularly important for FRED data (FRED provides a logarithmic key mapping).

**Frequency Alignment**: Transform the monthly economic indices from FRED to a quarterly format to align with the BEA’s quarterly GDP data. Calculate the sum or average (as appropriate) of monthly values within each quarter.

**Seasonal Adjustments**: Adjust high-frequency data for seasonality, if necessary, to isolate the core economic trends from regular seasonal patterns. This step will make the data more representative of general economic behaviours, irrespective of seasonal influences.

**Monthly Rate of Change**: For indices that are better represented through changes (e.g., stock indices, employment rates), calculate the month-over-month rate of change post-log transformation. This helps to highlight immediate shifts in economic activities.  
  
**Quarterly Integration**: Integrate the monthly indices into the quarterly GDP data framework. For BEA's GDP data, represent them as absolute figures or calculate the quarter-over-quarter rate of change if it aligns better with our analysis objectives.  
  
**Final Aggregation and Comparison**: Ensure that the final format of both datasets (quarterly GDP and monthly indices) is compatible for direct comparison. This could involve representing both datasets as rates of change or absolute figures based on what is most meaningful for the analysis.

4.3 Data Integration and Quality Assurance

**Data Integration**: We will merge various datasets into a unified framework using pandas, ensuring seamless integration and compatibility. This step is vital for consolidating different economic indicators into a single, comprehensive analysis.

**Final anomaly Detection and Correction**: Employing statistical methods to detect and correct anomalies ensures that our analysis is based on accurate and representative data, free from distortions that could lead to erroneous conclusions.

**Consistency Checks**: Conducting thorough checks for data consistency, especially when integrating diverse data sources, is essential to validate the reliability and accuracy of our findings.

4.4 Advanced Data Handling and Analysis

**Standardisation of Growth Rates**: Standardizing growth rates enables us to compare different economic indicators on a common scale, facilitating a more meaningful analysis across various data points.

**Stationarity Assessment**: Using tests like the Augmented Dickey-Fuller ensures that our time series data is suitable for modelling and forecasting, as many statistical models require stationarity for valid results.

**Addressing Non-Stationarity**: Techniques such as differencing or transformation will be applied to achieve stationarity, which is crucial for the accuracy and reliability of our predictive models and correlation analysis.

## 5. Analysis

5.1 Data Harmonization and Cleaning  
  
Conducted as the first step to create a reliable foundation for all subsequent analyses.

* **Technique**: see data cleaning and manipulation.
* **Objective**: To ensure both datasets (quarterly GDP and monthly economic indices) are in a compatible format and frequency for accurate comparative analysis.

5.2 Exploratory Data Analysis (EDA)

Performed early in the project to get an overview of the data's characteristics. This step is crucial for identifying the most relevant variables for analysis, understanding the data's basic structure, and ensuring that hypotheses are grounded in both statistical findings and economic logic.

* **Technique**: Using statistical tools to summarise the data, visualising distributions with histograms, identifying correlations with scatter plots, and detecting patterns and outliers with box plots.
* **Objective**: To gain an initial understanding of data trends, outliers, and correlations and to identify any anomalies or irregularities that may influence further analysis. Hereafter we will incorporate economic theories to hypothesise potential relationships between variables.

5.3 Seasonality Adjustment Analysis

Conducted post-EDA to refine the data for more accurate correlation analysis. Seasonality adjustment is essential for preventing seasonal patterns from distorting the true economic trends.

* **Technique**: Applying time-series decomposition methods to separate the data into trend, seasonal, and residual components and then adjusting for these seasonal effects.
* **Objective**: To accurately capture the underlying trends in consumer spending by removing repetitive seasonal patterns, which are regular but not necessarily related to the economic indicators of interest. *While crucial, we have to ensure this doesn't lead to an overly complex focus on time-series analysis techniques unless they are directly relevant to identifying proxies.*
  1. Correlation and Proxy Validation Analysis

Implemented after seasonality adjustments to ensure that the relationships being analysed and the proxies being identified are not influenced by seasonal fluctuations and confirm that identified relationships are economically plausible as well as statistically significant.

* **Technique**: Calculating Pearson or Spearman correlation coefficients to quantify the strength and direction of the relationship between different variables. Scatter plots will be used for a more nuanced view of these relationships.
* **Objective**: To identify which monthly indicators from the high-frequency dataset show a strong and statistically significant correlation with quarterly consumer spending figures. Economic theory will be applied to interpret these correlations, ensuring they align with established economic principles and behaviors.
  1. Comparative and Temporal Analysis

Undertaken after correlation analysis to delve deeper into the dynamics of the relationships between consumer spending and the identified proxies, providing insights into potential causative or predictive trends.

**Lead and Lag Analysis**:

* **Technique**: Analysing the time-shifted relationships between consumer spending and the proxies to identify if any indicators consistently lead or lag behind consumer spending patterns.
* **Objective**: To discover predictive relationships where certain proxies might signal changes in consumer spending ahead of time or respond with a delay. *While relevant, the Lead and Lag Analysis could become complex and time-consuming. We need to ensure that it directly contributes to the goal of identifying proxies.*

**Consumer Behaviour Indicators Correlation**:

* **Technique**: Using scatter plots and heatmaps to examine how different indicators relate to consumer spending visually.
* **Objective**: To explore more complex relationships between consumer spending and various high-frequency proxies and to identify patterns not evident in standard correlation analysis.

#### 5.6 Proxy Evaluation and Variable Selection

Essential for finalising the selection of proxies, ensuring they are representative of consumer spending trends and robust under different conditions.

**Variable Selection and Reduction**:

* **Technique**: Selecting proxies based on correlation outcomes and economic rationale.
* **Objective**: To focus on a select group of high-frequency proxies that most accurately reflect and predict trends in consumer spending.

#### Regression Analysis and Uncertainty Assessment

Performed as a concluding analytical step to provide a more nuanced understanding of how each identified proxy affects consumer spending. This step helps quantify the relationships discovered in earlier analyses.

**Model Evaluation and Uncertainty Assessment**:

* **Technique**: Utilizing advanced statistical techniques, such as bootstrapping or Monte Carlo simulations, to evaluate the robustness of the selected proxies.
* **Objective**: To assess the reliability and stability of the chosen proxies under various economic scenarios and conditions. *Techniques like bootstrapping or Monte Carlo simulations might be more advanced than required for this project as the primary aim is to identify proxies rather than to build a predictive model.*

**Regression Analysis**

* **Technique**: Conduct linear regression analysis to quantify the impact of each selected proxy on consumer spending and assess the significance of regression coefficients.
* **Objective**: To determine the strength and nature of the influence that each proxy has on consumer spending, thereby providing a quantitative measure of their relative importance.

## 6. Contributions

#### **Week 1: Data Preparation and Initial Analysis**

#### Data Collection and Initial Preparation

**TM 1**: Collect and organise the primary dataset from the U.S. Bureau of Economic Analysis.

**TM 2**: Collect and organise the secondary dataset from the Federal Reserve Economic Data (FRED).

**TM 3**: Begin setting up the data processing environment (Python, Pandas).

#### Data Cleaning and Standardization

**TM 1**: Handle missing values and anomalies in the BEA dataset.

**TM 2**: Handle missing values and anomalies in the FRED dataset.

**TM 3**: Assist with data cleaning as needed and ensure data types are standardized across datasets.

#### Data Harmonization

**All TMs**: Collaboratively work on log transformations, frequency alignment, and seasonal adjustments. Divide the tasks evenly, ensuring each member is aware of the methodologies and outcomes of the other datasets.

#### **Week 2: In-depth Analysis and Proxy Identification**

#### Exploratory Data Analysis (EDA)

**TM 1**: Focus on EDA for the BEA dataset.

**TM 2**: Focus on EDA for the FRED dataset.

**TM 3**: Start integrating findings from both datasets, identifying preliminary correlations.

#### Correlation and Proxy Validation Analysis

**TM 1 & 2**: Conduct correlation analysis on their respective datasets.

**TM 3**: Oversee the correlation analysis, ensuring consistency and accuracy.

#### Comparative and Temporal Analysis

**TM 3**: Lead the lead and lag analysis and consumer behaviour indicators correlation.

**TMs 1 & 2**: Provide support with additional analysis and data validation.

#### **Week 3: Finalization and Reporting**

#### Proxy Evaluation and Variable Selection

**All TMs**: Collaborate to select the most relevant proxies. Each member should evaluate a portion of the proxies and then discuss findings as a group.

#### Regression Analysis and Uncertainty Assessment

**TM 1**: Conduct regression analysis.

**TM 2 & 3**: Work on uncertainty assessment, possibly applying advanced statistical techniques.

#### Finalizing Report and Presentation

**All TMs**: Contribute to the final report. Assign sections to each member based on their primary areas of work during the project.

**Final Day**: Review and rehearse the presentation as a team. Ensure that each member is familiar with all aspects of the project.

#### **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.
* If any TM is ahead or behind, redistribute tasks to balance the workload.
* Keep a flexible approach to unexpected challenges, such as data discrepancies or anomalies.

#### 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 TMs.

#### 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 TM’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 our Colab notebooks using the requirements.txt file from our GitHub repository to install the same dependencies in Colab.

#### Data Analysis and Visualization

**Python:** The primary programming language for our data analysis.