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.

**Understanding and Preparing the Data**

**Initial Data Analysis**:

**Quarterly Data**: Start with the quarterly dataset to identify key economic indicators that closely align with consumer spending patterns.

**Monthly Data**: Examine the monthly dataset, focusing on variables that theoretically influence consumer spending (e.g., retail sales, employment rates).

**Harmonizing Data**:

**Time Frame Alignment**: Transform monthly data into a quarterly format for direct comparison. Aggregate monthly data (sum or average) to align with the quarterly periods.

**Normalization and Scaling**: Ensure consistent units and scales across datasets. Techniques like Min-Max Scaling or Z-score normalization may be applied.

**Correlation Analysis**:

Use Pearson or Spearman correlation coefficients to identify relationships between monthly and quarterly indicators.

Investigate lagged correlations, considering that consumer spending can be influenced by prior economic conditions.

**Refining Proxies for Nowcasting**

**Feature Selection**:

Narrow down to monthly indicators that mirror quarterly consumer spending trends.

Consider PCA or Factor Analysis for dimensionality reduction and identifying impactful features.

**Robustness Checks**:

**Historical Validation**: Test the proxies against historical data to evaluate their predictive accuracy.

**Cross-Validation**: Employ statistical methods to assess the stability of the findings.

**Additional Methods and Considerations**

**Economic Theory Integration**: Ensure the chosen indicators are grounded in economic theory and logic.

**Leading Indicator Analysis**: Identify if any monthly indicators are leading indicators, providing early signals for consumer spending changes.

**Use of External Data Sources**: Consider including additional data sources like consumer sentiment indices, stock market indices, or high-frequency data like credit card transaction volumes.

**Computational Narrative**: Document each analysis step, explaining the rationale and findings. Utilize Python libraries (Pandas, Matplotlib, Seaborn, Scikit-learn) for data manipulation and visualization.

**Approach for Using Proxies**

**Data Collection**: Aggregate historical data for all selected indicators.

**Detailed Correlation Analysis**: Determine the strongest historical correlations with consumer spending.

**PCA or Factor Analysis**: Use these techniques to identify principal components explaining the most variance in consumer spending.

**Model Selection**: While you're not building a model, suggest appropriate econometric or machine learning models that could utilize these proxies for prediction.

**Feature Engineering**: Consider using lagged values, moving averages, or percentage changes of indicators to capture underlying trends.

**Short outline:  
Framework for Analyzing Consumer Expenditure:**

**1. Consumer Spending and Income:**

Real Personal Consumption Expenditures: Measure spending on durable goods, nondurable goods, and services.

Real Personal Income: Total and excluding transfer receipts, providing a broad view of income.

Real Disposable Personal Income: Post-tax income directly drives spending capacity.

Retail and Food Services Sales: Immediate reflection of consumer spending trends.

**2. Consumer Sentiment and Confidence:**

Consumer Sentiment Index: Directly assesses consumer attitudes and willingness to spend.

Economic Policy Uncertainty Index: Offers context on how external economic factors might impact consumer confidence and spending.

**3. Employment and Labor Market:**

All Employees Data (Various Sectors): Connects employment levels to spending potential across sectors.

Civilian Labor Force Statistics: Encompasses employment, unemployment rates, and duration of unemployment.

Average Weekly Hours and Overtime: Indicates labor market tightness and income stability.

**4. Industrial Production and Business Activity:**

Industrial Production Index: Highlights overall industrial activity that feeds into consumer spending.

ISM Manufacturing Indexes (e.g., Production, New Orders, Employment): Manufacturing activity as a precursor to consumer demand.

Capacity Utilization: Signaling production adjustments to meet consumer demand.

**5. Housing Market:**

Housing Starts and Building Permits: Signifies consumer investment levels and confidence, which typically lead to related consumer spending.

**6. Inflation and Price Indices:**

Consumer Price Index (CPI): Measures consumer inflation, directly impacting purchasing power.

Producer Price Index (PPI): Preemptively indicates potential consumer price changes.

**7. Monetary and Financial Indicators:**

M1 and M2 Money Stock: Money supply measures that can predict inflationary pressures.

Real Estate and Commercial and Industrial Loans: Credit levels can forecast future spending and investment.

**8. Interest Rates and Financial Market Conditions:**

Federal Funds Rate, Treasury Rates, Corporate Bond Yields: The cost of borrowing and the yield environment, which influences spending and saving decisions.

S&P’s Stock Price Indexes: The wealth effect from stock market valuations can affect consumer spending.

**9. Exchange Rates and International Trade:**

Trade Weighted U.S. Dollar Index: Impacts the cost of imported goods, influencing consumer prices and spending.

Foreign Exchange Rates: May affect the competitiveness of domestic industries and consumer prices.

Approach for Using Proxies to Predict or Nowcast Consumer Spending:

* Data Collection: Aggregate historical data for all selected indicators.
* Correlation Analysis: Determine which indicators have the strongest historical correlation with consumer spending.
* PCA or Factor Analysis: Identify the principal components that explain the most variance in consumer spending.
* Model Selection: Choose appropriate econometric or machine learning models for prediction.
* Feature Engineering: Use lagged values, moving averages, or percentage changes of indicators to capture trends.
* Sensitivity Analysis: Assess how sensitive the model is to changes in each indicator.

## 

## 1.1Introduction: The Problem (Why)

GDP reports are slow to capture economic changes, which may affect multiple stakeholders. To address this, a project aims to provide more immediate insights into consumer behavior and spending trends by using high-frequency data proxies for consumer expenditure.

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

The primary objective is to find and confirm alternative proxies that can be used to nowcast consumer expenditure in the United States. This will provide an up-to-date snapshot of consumer spending habits. The main challenge is to overcome the delay in consumer spending data reporting and to provide relevant economic indicators in real-time. To achieve this, we will use a systematic approach that involves data analysis, proxy validation, and integration. This will enable us to produce an accurate and immediate 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?

#### 1.3 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.

#### 1.4 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.

#### 1.5 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.

#### 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. 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.

**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 Secondary Datasets

#### 3.1 Dataset 1: Federal Reserve Economic Data (FRED)

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 both monthly intervals, with a particular focus on US GDP data. The data covers consumer spending indicators, which are a crucial component of the Gross Domestic Product (GDP). 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.

* **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.

#### 3.2 E-commerce Sales Data to be determined

* **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 Google trends

To be determined

## 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?

#### 4.1 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.

#### 4.2 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.

#### 4.3 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.

#### 4.4 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.

#### 4.5 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.

#### 4.6 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.

## 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.

#### 5.1. 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.

#### 5.2. 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.

#### 5.3. 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.

#### 5.4. 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.

#### 5.5. 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.

#### 5.6. 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.

## Contributions

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

#### **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.