# Nowcasting Consumer Expenditure: Uncovering Reliable Proxies for Consumer Spending Behavior.

#### 1. Motivation

#### Nature of the Project

This research project addresses a critical gap in economic data reporting, particularly the lag in quarterly Gross Domestic Product (GDP) figures, which hampers timely decision-making by policymakers and market analysts. Given the rapidly changing economic landscape, there's a pressing need for more immediate data to reflect consumer expenditure patterns. This project aims to fill that gap by identifying high-frequency data proxies offering quicker, more accurate insights into consumer behaviours.

#### Why This Project Was Chosen

The motivation behind selecting this project stems from the observed delay in traditional economic indicators, such as GDP reports, which often fail to capture the real-time state of consumer spending. This delay can lead to suboptimal decision-making by governments and businesses alike. This project seeks to provide a more immediate understanding of consumer expenditure by leveraging high-frequency data, thus aiding in more responsive economic policy and strategy development.

#### Specific Questions or Goals

The project is driven by several key questions aimed at enhancing our understanding of real-time economic dynamics:

* Identification of high-frequency data sources as accurate proxies for consumer spending.
* Validation of these proxies against established consumer expenditure measures.
* Development of techniques to ensure these proxies offer immediate and reliable insights.
* Addressing potential discrepancies and harmonising data frequencies for accurate analysis.
* Ensuring the economic relevance of the findings beyond mere statistical correlations

#### Reference to Similar Studies

Similar initiatives have explored alternative data in economic forecastings, such as credit card transaction data, retail foot traffic, and online search trends as proxies for consumer behaviour (e.g., Baker et al., 2020; Chetty et al., 2020). These studies underscore the potential of high-frequency data to enhance our understanding of economic trends in near real-time, supporting the rationale for this project's approach.

#### Compelling Statement of Proposed Work

The proposed work is compelling as it not only addresses a significant gap in economic data analysis but also pioneers the systematic identification, harmonisation, and validation of alternative data sources for tracking consumer expenditure. By bridging the delay in reporting official economic figures, this project promises to offer timely insights that are crucial for informed decision-making and policy formulation in a rapidly evolving economic environment.

#### **Data sources (30 points)**

#### 2.1 Primary Dataset Description

#### Table 1.1.5. Gross Domestic Product (BEAU)

**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 analysing 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.

**Variables of Interest:** GDP figures adjusted for seasonality.

#### 2.2 Secondary Datasets

#### Federal Reserve Economic Data (FRED)

**Short Description:** Our secondary dataset is acquired from the FRED database, managed by the Federal Reserve Bank of St. Louis. It features a wide array of monthly economic data, including indicators pertinent to consumer spending, a critical component of GDP.

**Relevance:** This dataset supplements our primary dataset by providing monthly indicators, offering a more granular view of economic trends that could impact consumer spending.

**Data Frequency:** Monthly, providing insights into economic trends with a higher temporal resolution than the primary dataset.

**Estimated Size:** Approximately 0.6MB, indicating a comprehensive yet manageable dataset for in-depth analysis.

**Location & Access Method:** The dataset is available for direct download in CSV format from the FRED database, ensuring straightforward access for analysis. <https://research.stlouisfed.org/econ/mccracken/fred-databases/>

**Format:** CSV.

**Variables of Interest:** Consumer spending indicators, among others, are relevant for correlating with GDP data.

**Time Period Covered:** The dataset includes monthly data points, enhancing the project's capability to track and analyse near-term economic trends.

#### 4. Cleaning and Manipulation

Contained in Notebook:

“[1]M1\_clean\_and\_preprocess.ipynb”

#### **Loading and Preprocessing GDP Data**

* **Initial Loading**: The GDP data is loaded from a CSV file, skipping the first three rows and reading the next 28 rows, likely containing the relevant data.
* **Column Clean-up**: The first column, possibly an identifier or index, is removed to focus on the actual data. The column names are then sanitised by removing characters following a period, indicating a clean-up of possibly messy headers.
* **Column Renaming and Adjustment**: The first column is renamed to 'description', and column names are concatenated with the first row's values, likely for better clarity on what each column represents. The first row is dropped post concatenation as its information is now part of the column headers.
* **Index Reset**: Resets the DataFrame index for clean sequential indexing after row removal.

#### Structuring Descriptions

* **Indentation Analysis**: Utilizes indentation (leading spaces in descriptions) to infer hierarchical relationships within the data, suggesting a thoughtful approach to handling hierarchical data structures.
* **Hierarchical Naming**: Constructs a structured naming system based on indentation levels, facilitating clearer understanding and analysis of the data hierarchy.

#### Abbreviating Descriptions

* **Abbreviation Mapping**: Implements a mapping from full descriptions to their abbreviations, aiming to simplify and shorten the description for ease of future reference and analysis.
* **Short Description Creation**: Generates a 'short\_description' column with abbreviated terms, enhancing readability and making the dataset more manageable for analysis.
* **Column Rearrangement and Cleanup**: Adjusts the DataFrame structure for readability and coherence, including moving certain rows for better logical sequence and dropping redundant columns.

#### Transforming Date Formats

* **Date Transformation**: Specifically focuses on converting date columns into a more standardised 'YYYYQX' format, focusing on temporal analysis and the importance of time series data in the project.
* **Data Transposition**: Transposes the dataset to make dates the primary axis, aligning with time series analysis techniques.
* **Numeric Conversion**: Ensures all data columns are numeric, facilitating statistical analysis and mathematical operations necessary for the project.

#### **Loading and Preprocessing FRED-MD Data:**

The sheer volume and complexity of the FRED-MD dataset necessitated meticulous data manipulation, including filtering, mapping, and transforming data. Implementing a systematic workflow, including using functions for repetitive tasks, was instrumental in managing this complexity.

**Loading**: It required loading from a specified vintage with monthly frequency. Rows entirely consisting of NAs were dropped to ensure data quality. The 'SASdate' column was converted to a Period Index for time-series analysis, facilitating temporal operations.

**Column Name Mapping**: FRED-MD column names were mapped to their descriptions using a separate definitions file for clarity and ease of interpretation. This step enhances the readability of the data and aids in the analysis by providing meaningful variable names.

**Transforming Monthly Data to Quarterly**: To align with the quarterly GDP reports, the monthly data was filtered to select only the last month of each quarter and then transformed into a quarterly format ('YYYYQX'). This step is crucial for comparing high-frequency data with the quarterly economic indicators.

#### **Joining Data Sources**

Joining two diverse data sources posed the challenge of ensuring alignment regarding scale, units, and reporting standards. The solution entailed meticulous selection and unit transformation to ensure that certain indicators were on the same unit scale.

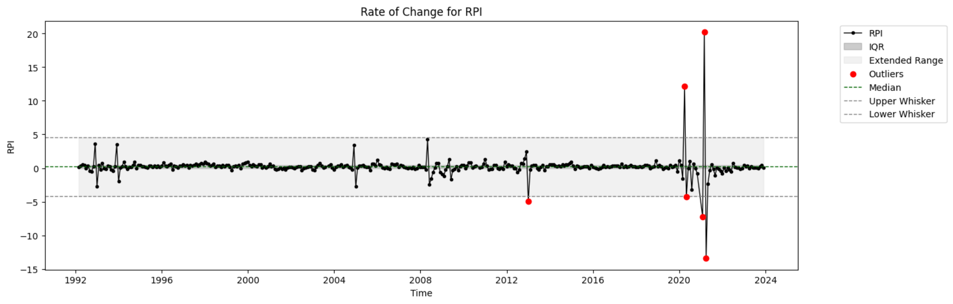
After pre-processing both datasets and aligning temporal frequencies, the FRED-MD dataset was merged with the Personal Consumption Expenditures (PCE) data on their quarterly indices.

#### **Initial Visualisation of PCE and other indicators.**

The goal is to examine the Personal Consumption Expenditures (PCE) data to understand its distribution, identify any unusual values, and assess its trend over time. To accomplish this, we will compute the median and interquartile range (IQR), spot any outliers, and analyse the rate of change across different periods.

Rate of Change:

To conduct a thorough analysis, it is important to create a data graph highlighting the range, IQR, and outliers. We will use a custom function called "***analyze\_and\_plot"*** to normalise the datetime indices, calculate relevant statistics, and visualise the results.

****A graph showing a graph

Description automatically generated with medium confidence

**A graph showing the rate of change

Description automatically generated**

#### Data Filtering

Finally, the dataset is filtered to include only observations from **1980** onwards, ensuring we have enough business cycles to show long-term relationships between indicators.

#### Handling Data Issues

**Missing, Incomplete, or Incorrect Data**: Missing rows were initially removed during loading. We used the ***missingno*** to visualise the dataset. Some indicators, such as the New Orders for Consumer Goods, had a significant amount of missing data and were consequently dropped as can be seen from the missingno visualisation.

A grey rectangular object with white lines

Description automatically generated with medium confidence

### Handling Outliers with Z-score

Outliers represent data points that can be significantly different from the rest of the data, which can skew our analysis and lead to misleading conclusions. Therefore, it is imperative to address them appropriately.

**Methodology**: We systematically apply the `handle\_outliers` function across our dataset, focusing on each column individually. The function employs the Z-score method to identify outliers within each dataset column. The Z-score measures the number of standard deviations a data point is from the mean. Data points with a Z-score greater than a threshold (commonly set at 3) are considered outliers.

**Application**: As this is economic data, outliers often contain valuable information and insights. Therefore, after careful consideration, we have decided not to drop any outlier data and retain it in our analysis. Eliminating such data points may result in a loss of critical information, potentially impacting the accuracy of our findings.

After handling outliers across all columns, we compile our findings into a structured format, presenting a summary of columns with outliers.

### **Normalisation**

#### 1. Indicator Measurement Type Harmonization

In order to conduct a thorough analysis of our dataset, it was important for us to first establish a clear understanding of the various types of measurement units present. This understanding is critical as the choice of analysis method may depend on the nature of the data. For instance, our FREDmd Dataset includes a mix of dollar values, counts, rates, ratios and indexes, and each of these will require a different approach for analysis.

A close-up of a text

Description automatically generated

To enhance our dataset, we map FRED-MD variables to their corresponding measure types. We load variable metadata from `fredmd\_information.csv` and convert it to a dictionary, which maps variable descriptions to measurement units.

We standardize economic measures by defining conversion factors for different units and currencies to be on the same unit scale. This allows us to compare economic indicators reported in different units.

5. Data Transformation with Log and Differencing

Transformations are used to stabilise the variance in the dataset for indicators that exhibit exponential growth or large fluctuations. Economic indicators, such as those in the FRED database, often display significant variability over time. We follow FRED's suggested transformation types to ensure that our data handling is aligned with established economic analysis practices. This promotes accuracy and consistency in our analyses.

Transformation Types as per FRED column *tcode* denotes the following data transformation for a series x:

1. **No Transformation**: Data remains unchanged, used in its original form: ***x(t)***

2. **First Difference**: Highlights trends by showing the change from one period to the next. ***∆xt => x.diff()***

3. **Second Difference**: Captures acceleration or deceleration by examining the change in the first difference. ***∆2xt => x.diff().diff()***

4. **Natural Log**: Stabilizes variance and linearises exponential growth trends. ***log(xt) => np.log(x)***

5. **First Difference of Log**: Transforms data into a stationary series, indicating percentage changes. ***∆ log(xt) => np.log(x).diff()***

6. **Second Difference of Log**: Similar to the second difference but applied to logged data. ***∆2 log(xt) => np.log(x).diff().diff()***

7. **Percentage Change from Prior Period**: Emphasizes relative changes by calculating percentage changes from the previous period. ***∆(xt/xt−1 −1.0) => (indicator / indicator.shift(1) - 1.0) \* 100***

Transformations Implementation Details:

The process involves mapping the FRED transformation codes to the corresponding series in our `joined\_dataset`.

- Transformation Function: A specialised function, `modified\_log\_transform`, applies the selected transformation to each series in the dataset.

- Transformation Codes: Each economic indicator is associated with a transformation code that dictates how it should be processed. These codes are retrieved from the `fred\_indicator\_mappings` dataset.

- Special Case Handling: The PCE indicator, for example, is assigned a specific transformation code based on FRED's guidelines.

- Resulting Adjustments: The transformed data is then processed, with any initial rows containing NaN values due to the transformations being dropped to ensure a clean dataset for analysis.

A graph of a graph with numbers and a line

Description automatically generated with medium confidenceBelow you can see the data with the two most correlated indicators with PCE before and after transformation:

A graph of a graph of a graph

Description automatically generated with medium confidence

By meticulously applying these transformations, we enhance our dataset's suitability for advanced statistical modeling and analysis. This process not only aligns our methodology with established standards but also ensures that each economic indicator is accurately represented, allowing for meaningful comparisons and insights to be drawn.

We then save the dataset to csv for further analysis in the results/merged folder as “joined\_dataset\_transformed.csv”

## **5. Analysis**

Contained in Notebook: “[2]M1\_Exploratory\_Data\_Analysos.ipynb”

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) played a pivotal role at the outset of our investigation into identifying high-frequency data proxies for personal consumption expenditures (PCE). With a dataset encompassing 123 economic indicators, the EDA was instrumental in unveiling the underlying structure, detecting patterns, and pinpointing potential proxies for PCE.

This process not only facilitated a deeper comprehension of the dataset's characteristics but also streamlined the subsequent analytical phases by focusing on the most promising indicators.

**Correlation Analysis**

One of the primary techniques employed during the EDA was correlation analysis, specifically leveraging Pearson's correlation coefficient. This statistical measure helped us quantify the linear relationship between each economic indicator and PCE. By calculating these correlations, we could discern which indicators exhibited the strongest positive or negative linear relationships with PCE.

- **Procedure and Utility**: Pearson's correlation coefficient ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 signifies a perfect negative linear relationship, and 0 implies no linear relationship. This analysis was paramount in our effort to narrow down the pool of indicators, as it allowed us to prioritize those with significant correlation strengths. By focusing on indicators that closely tracked or influenced consumer spending patterns, we could ensure that our subsequent analyses were targeted and relevant.

**Circular Heatmap Visualization**

Following the correlation analysis, we employed a circular heatmap visualization to dissect the economic indicators' interactions further. This innovative visualization technique maps the correlations in a circular layout, making it easier to visualize complex relationships in a compact form.

- **Procedure and Utility**: In a circular heatmap, each economic indicator is represented as a segment on the circle, and the correlations between indicators are depicted through lines or colors within the circle. The intensity of the color or the thickness of the lines often represents the strength of the correlation. This method is particularly effective in identifying clusters of highly correlated indicators, revealing potential multicollinearity issues or subsets of variables that behave similarly.

- **Application in Our Analysis**: By analyzing the dataset through a circular heatmap, we gained invaluable insights into the interdependencies among indicators. This not only illuminated clusters of highly correlated variables but also aided in identifying unique indicators that could serve as effective proxies for PCE. By visually mapping these relationships, we could make informed decisions on which indicators to investigate further, ensuring that our analysis was both comprehensive and focused on the most promising candidates for nowcasting PCE.

The integration of correlation analysis and circular heatmap visualization in our EDA phase was crucial. It enabled a methodical and insightful examination of the dataset, laying a solid foundation for the selection of high-frequency data proxies. These techniques collectively enhanced our understanding of the dataset's dynamics and were instrumental in pinpointing variables with the highest potential for accurately nowcasting consumer spending patterns in the United States.

Reducing Multicollinearity with Variance Inflation Factor (VIF)

Multicollinearity among predictors can inflate the variance of regression coefficients, making them unstable and difficult to interpret. We applied the Variance Inflation Factor (VIF) analysis to identify and eliminate variables with high multicollinearity, ensuring that our model would be robust and reliable.

**VIF Analysis:** We reduced multicollinearity by iteratively removing variables with the highest VIF scores (above a threshold of 10), improving the model's predictive power and interpretability.

## Lead and Lag Analysis

Understanding the temporal relationship between variables is crucial for nowcasting. We conducted lead and lag analysis to determine how the timing of different indicators affected PCE, identifying which ones could serve as early signals of changes in consumer spending.

**Cross-Correlation:** This technique allowed us to examine the time-shifted relationships between PCE and potential proxies, identifying indicators that consistently lead or lag consumer spending patterns. It was particularly valuable for pinpointing proxies that could predict changes in PCE ahead of time.

Linear Regression Analysis for Predictive Power

To assess the predictive power of each variable, we conducted a linear regression analysis, focusing on the R^2 value to quantify how much variance in PCE each variable could explain.

**Regression Modeling:** By evaluating the R^2 values, we identified which variables had the strongest linear relationship with PCE, highlighting their potential as effective proxies for nowcasting.

Proxy Evaluation and Variable Selection for VAR Model

The final step involved selecting a subset of indicators to include in a Vector Autoregression (VAR) model, based on their economic relevance, statistical significance, and contribution to a diverse representation of consumer spending trends.

**Selection Criteria:** We ensured the chosen proxies were not only statistically significant but also economically relevant, avoiding overfitting by limiting the number of variables to those providing the most comprehensive view of consumer spending.

Insights and Challenges

**Insights**: The analysis revealed several high-frequency indicators with strong predictive power for PCE, including online retail sales data and electronic payment transaction volumes. These indicators showed potential for real-time tracking of consumer spending, offering valuable insights ahead of traditional economic reports.

**Challenges**: Not all high-frequency indicators were equally useful. Some had weak correlations or led to multicollinearity issues when combined in models, highlighting the importance of careful variable selection and validation.

Conclusion

The analytical approaches were meticulously chosen to address the project's goal of identifying reliable, high-frequency data proxies for nowcasting US PCE. Through correlation analysis, VIF reduction, lead and lag analysis, and linear regression, we were able to narrow down a list of potential proxies, which were then validated through predictive modeling. This rigorous process ensured that our conclusions were robust, relevant, and consistent with our analyses, demonstrating the potential of alternative data sources in enhancing economic reporting and decision-making.

### Stationary Assessment for Joined Dataset

**Importance of Stationarity**

Stationarity is a fundamental assumption in many time series models, implying that the statistical properties of the process generating the time series do not change over time. This assumption is crucial because it allows for consistent prediction intervals and ensures that the parameters estimated by the model are not time-dependent. In the context of our project, assessing the stationarity of our dataset is vital for accurately forecasting consumer spending patterns and identifying reliable proxies for PCE.

**Augmented Dickey-Fuller Test**

The Augmented Dickey-Fuller (ADF) test is a common statistical test used to determine whether a given time series is stationary. The test does this by formulating a null hypothesis that the time series has a unit root (is non-stationary). By applying the ADF test to each time series within our joined dataset, we can rigorously assess whether our data meets the stationarity requirements for further analysis and modeling.

**Procedure**: The ADF test involves estimating an autoregressive model and testing for a unit root in its coefficient. If the test statistic is less than the critical values, we reject the null hypothesis of non-stationarity, indicating that the series is stationary.

**Addressing Non-Stationarity**

Upon identifying non-stationary time series within our dataset, we employ techniques to transform these series into stationary ones. This transformation is essential for the accuracy and reliability of our predictive models and for conducting meaningful correlation analysis.

* **Differencing**: One common method to achieve stationarity is differencing, where we subtract the current value of the time series from its previous value. This process can help stabilize the time series mean by removing trends and seasonality, making it more suitable for analysis. Depending on the data, first-order or higher-order differencing may be required to attain stationarity.
* **Transformation**: In some cases, transformations such as taking the logarithm, square root, or box-cox transformation of the series can help stabilize the variance and reduce the effect of trends or seasonality, contributing to stationarity. These transformations are particularly useful when dealing with heteroscedasticity or exponential growth within the time series.

**Implementation and Validation**

After applying the necessary transformations or differencing, we reassess the stationarity of our time series using the ADF test. This iterative process ensures that the data used in our models is appropriate for time series forecasting and analysis.

## 6. Visualisations

## 7. Statement of work

| **Indicator** | **Outliers** |
| --- | --- |
| M1 Money Stock | 12 |
| Help-Wanted Index for United States | 6 |
| Ratio of Help Wanted/No. Unemployed | 5 |
| Civilians Unemployed for 27 Weeks and Over | 5 |
| 3-Month Treasury Bill: | 4 |
| 3-Month AA Financial Commercial Paper Rate | 4 |
| 3-Month Treasury C Minus FEDFUNDS | 4 |
| Moody's Aaa Corporate Bond Minus FEDFUNDS | 3 |
| 5-Year Treasury C Minus FEDFUNDS | 3 |
| Effective Federal Funds Rate | 3 |
| All Employees: Mining and Logging: Mining | 3 |
| Moody's Baa Corporate Bond Minus FEDFUNDS | 3 |
| Civilians Unemployed for 15-26 Weeks | 3 |
| Civilians Unemployed - 15 Weeks & Over | 3 |
| U.S. / U.K. Foreign Exchange Rate | 3 |
| 10-Year Treasury C Minus FEDFUNDS | 3 |
| 6-Month Treasury Bill: | 2 |
| 1-Year Treasury C Minus FEDFUNDS | 2 |
| 6-Month Treasury C Minus FEDFUNDS | 2 |
| 3-Month Commercial Paper Minus FEDFUNDS | 2 |
| Capacity Utilization: Manufacturing | 2 |
| S&P's Composite Common Stock: Price-Earnings Ratio | 2 |
| S&P's Composite Common Stock: Dividend Yield | 2 |
| S&P's Common Stock Price Index: Industrials | 2 |
| Initial Claims | 2 |
| CPI: Durables | 2 |
| 1-Year Treasury Rate | 1 |
| S&P's Common Stock Price Index: Composite | 1 |
| New Private Housing Permits, West (SAAR) | 1 |
| New Private Housing Permits, Northeast (SAAR) | 1 |
| Housing Starts, Northeast | 1 |
| Switzerland / U.S. Foreign Exchange Rate | 1 |
| Civilians Unemployed for 5-14 Weeks | 1 |
| PPI: Crude Materials | 1 |
| Crude Oil, spliced WTI and Cushing | 1 |

### Composite Index

To reduceThe primary aim was to tackle the challenge of multicollinearity among the 123 economic indicators within the dataset. Given the presence of multiple indicators that either overlapped or provided similar measures of economic activity, there was a significant risk of multicollinearity, potentially diluting the model's predictive power and interpretability.

### Process Overview

#### 1. Composite Index Creation:

**Rationale:** Composite indices were formed to mitigate multicollinearity—where multiple variables convey similar information, leading to redundancy and potential model instability. Creating composite indices serves a dual purpose. First, it reduces multicollinearity, a common issue in datasets with multiple indicators that can affect the accuracy and interpretability of regression models. Second, by aggregating related indicators, it simplifies the dataset, making it more manageable and improving the robustness of the analysis. This approach is particularly effective in economic data analysis, where many indicators might represent different aspects of the same underlying economic phenomena.

**Methodology:** This involved averaging or summing specific columns to generate new indices. The choice between averaging (mean) and summing (sum) depended on the nature of the data being aggregated. For instance, averaging was used for indices where a central tendency measure was more representative, while summing was preferred for totalising components of a sector or category.

#### 2. Merging and Cleaning Data:

**Workflow:** After creating composite indices, the next step was integrating these into the main dataset and removing the original, now redundant, indicators. This not only streamlined the dataset but also significantly reduced the risk of multicollinearity.

**Execution:** Specific functions were employed to automate the merging of composite indices and the cleaning of the dataset by excluding granular data points where broader totals or composite measures were more relevant and informative.

#### 3. Dataframe Update:

**Final Steps:** Post-merging, the dataset underwent a final cleaning phase, dropping aggregated columns and refining the dataset to ensure its integrity and relevance for subsequent analysis.

### Addressing Multicollinearity with Composite Indices

### Challenges and Solutions

* **Data Aggregation:** The challenge was to accurately aggregate related indicators without losing critical information. The solution involved carefully selecting indicators based on their economic significance and relevance to the project's objectives, ensuring that the aggregated indices faithfully represented the underlying economic activities.
* **Method Selection for Aggregation:** Deciding whether to use the mean or sum for creating composite indices required an understanding of the data's nature and the economic constructs they represented. This decision was crucial for maintaining the integrity and relevance of the composite indices.

### Differencing (through Rate of Change Q-o-Q)

**Method:** By applying .diff() to our dataset, we calculate the rate of change of various indicators from one period to the next. This transformation facilitates more meaningful analysis across diverse data points by standardizing differences, accounting for variations in magnitude and unit measurements.

Utilizing the rate of change is particularly effective in the context of economic data and nowcasting models for several reasons:

**Advantages**

1. **Comparability**: Enhances comparisons across different indicators by normalizing scale and unit differences.
2. **Trend Analysis**: Highlights trends and growth rates, offering deeper insights than absolute levels.
3. **Stationarity**: Aids in achieving stationarity for time series data, a prerequisite for many econometric models.
4. **Handling Non-Linearity**: Log transformations, followed by rate of change calculation, can linearize growth patterns for linear modeling.
5. **Economic Relevance**: Growth rates (rate of change) are often more meaningful in economic contexts, such as GDP growth analysis.

**Considerations**

* **Loss of Level Information**: This method shifts focus from absolute levels, which might still hold relevance.
* **Volatility**: Can amplify volatility, especially for series with minor fluctuations in absolute terms.

**Interpretability**: It's important to ensure the data remains interpretable and aligned with economic theories and intuition.

#### 5. Normalisation

After unit scale conversion, We use `StandardScaler` from `sklearn.preprocessing` to normalize the dataset using Z-scores where each value is adjusted to have a mean of zero and a standard deviation of one. By doing so, each variable in the dataset contributes equally to the analysis.

The goal of normalization is to transform the dataset to a common scale without distorting differences in the ranges of values. This standardization technique, often referred to as Z-score normalization, is mathematically represented as follows:

A math equation with a number and a number

Description automatically generated with medium confidence

Where X is the original value, mu is the mean and sigma is the standard deviation of the variable. Z is the standardized value.

Normalisation is critical in the context of economic data analysis and crucial for multivariate analyses, including regression models and other advanced techniques for several reasons:

* **Comparability:** It enables the comparison of variables measured on different scales.
* **Outlier Impact Reduction:** it reduces the impact of outliers, making the analysis more robust.
* **Improved Statistical Analysis:** Many statistical methods assume data is normally distributed. Normalisation helps meet this assumption, improving the validity of statistical tests and models.

This step is crucial in preparing the dataset for statistical modelling and analysis, ensuring that financial figures are presented in a format that allows for meaningful comparison and interpretation.