Nowcasting Consumer Expenditure:

Uncovering Reliable Proxies for Consumer Spending Behavior.

Why This Project Was Chosen

This project tackles a significant shortfall in economic data analysis, focusing on the delay in reporting quarterly Gross Domestic Product (GDP) figures, which adversely affects the ability of policymakers and market analysts to make timely decisions.

In the context of a rapidly evolving economic environment, the demand for immediate data reflecting consumer spending patterns is increasingly urgent. Our aim is to bridge this gap by identifying high-frequency data proxies that provide faster and more precise insights into consumer behaviour.

The imperative for this project is rooted in the notable lag associated with traditional economic indicators like GDP reports, which are often inadequate in capturing the instantaneous state of consumer spending. Such delays can lead to sub-optimal decision-making processes among governments and businesses. By harnessing high-frequency data, this project endeavours to furnish a more immediate comprehension of consumer expenditure, thereby supporting the formulation of more agile economic policies and strategies.

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 forecasting, such as credit card transaction data, retail foot traffic, and online search trends as proxies for consumer behaviour. 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. We found specific inspiration from this research done by *McCracken, M.W., Ng, S., 2015; FRED-MD: A Monthly Database for Macroeconomic Research, Federal Reserve Bank of St. Louis Working Paper 2015-012. URL https://doi.org/10.20955/wp.2015.012*

Compelling Statement of Proposed Work

The proposed work is compelling as it addresses a significant gap in economic data analysis and 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 crucial for informed decision-making and policy formulation in a rapidly evolving economic environment.

Key Audiences

- Policymakers and Government Officials: Benefit from real-time insights for responsive economic decision-making.
- Economic Analysts and Researchers: Interested in advanced economic forecasting and analysis methods.
- Financial Institutions and Market Analysts: Seek immediate data for informed investment strategies.
- Business Leaders and Strategists: Require up-to-date consumer behavior insights for strategic planning.

Team: Jan Nagtegaal | Aditya Sharma | Evan Trout

GitHub repository: https://github.com/janjannagtegaal/NowCasting-M1

Nowcasting Consumer Expenditure:

	Gross Domestic Product (BEAU)	Federal Reserve Economic Data (FRED)
Short Description:	The dataset "Table 1.1.5. Gross Domestic Product" from the U.S. Bureau of Economic Analysis comprises seasonally adjusted quarterly U.S. Gross Domestic Product (GDP) rates and its components in billions of dollars.	The FRED database is managed by the Federal Reserve Bank of St. Louis and features 123 monthly economic indicators.
Relevance:	The US GDP dataset's detailed information over several years is crucial for nowcasting consumption. Its granularity and time-series nature allow for comprehensive analysis and trend identification, making it pivotal for project success.	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:	The data reflecting the economic output of the United States is is done quarterly by the GDP component.	Monthly, providing insights into economic trends with a higher temporal resolution than the primary dataset.
Location:	Available at U.S. Bureau of Economic Analysis. (BEA)	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 Approximately 0.4 MB	CSV Approximately 0.6MB,
Access Method:	The dataset is readily available and can be easily accessed and downloaded directly from the U.S. Bureau of Economic Analysis website.	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:	The target indicator we are interested in for Nowcasting is Private Consumption Expenditure. (PCE)	The indicators collected from various economic sectors present a great dataset that we can use to evaluate and identify alternative proxies for Nowcasting.

Loading and Preprocessing GDP Data

- Initial Loading: The GDP data is loaded from a CSV file, skipping the first three title and summary rows and reading the next 28 rows.
- Column Clean-up: The first column an index, is removed to focus on the actual data. We then also remove all leading and trailing spaces.
- 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.
- Index Reset: Resets the DataFrame index for clean sequential indexing after row removal.

Structuring Descriptions

- Hierarchical Naming: Constructs a structured naming system based on indentation levels as the CSV contains indentation to show component hierarchies.
- Abbreviation Mapping: We implement a mapping from full component descriptions to their abbreviations, aiming to simplify and shorten the description for ease of future reference and analysis.
- Short Description Function: Generates a 'short_description' column with abbreviated terms, and we extract only the PCE series and its components.

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:

- Loading: Retrieving the latest version `current` of the FRED-MD dataset based on the specified `vintage`, ensuring our analysis is grounded on the most current data. 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').

Joining Data Sources

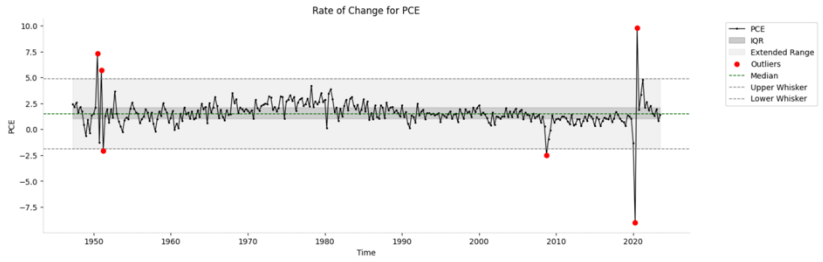
After pre-processing both datasets and aligning temporal frequencies, the FRED-MD dataset was merged with the Personal Consumption Expenditures (PCE) data on their quarterly indices and assigned to joined_dataset.

Initial Visualisation of PCE and other indicators.

Initial data inspection: The next step is to examine the Personal Consumption Expenditures (PCE) data to understand its distribution, identify any unusual values, and assess its trend over time. We did this by calculating the rate of change of PCE:

$$\left(\frac{(Current\ Value - Previous\ Value)}{Previous} Value\right) X\ 100\%$$

Creating a data graph highlighting the range, IQR, median and outliers is important to conduct a thorough analysis. We will use a custom function called "*analyze_and_plot*" to normalise the datetime indices, calculate relevant statistics, and visualise the results.



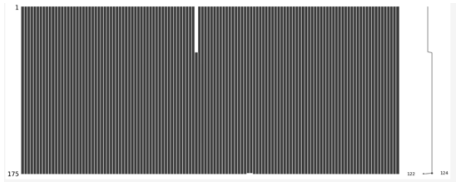
Data Filtering

Date Range 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.

Column filtering: We are also removing certain columns that are deemed less reliable for macro-economic research, such as the "Help-Wanted Index for United States" and the "Ratio of Help Wanted/No. Unemployed" due to their reliability concerns. Additionally, I am dropping the "S&P's Composite Common Stock: Price-Earnings Ratio" because it is typically reported with a significant six-month delay, and the "Consumer Sentiment Index," which is not only available quarterly prior to November 1977 but also suffers from a one-year lag in recent data availability in the FRED database as highlighted by McCracken in their 2015 working paper for the Federal Reserve Bank of St. Louis

Handling Data Issues

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 seen from the missingno visualisation.



Handling Outliers with Z-score

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. Data points with a Z-score greater than a threshold (commonly set at 3) were 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

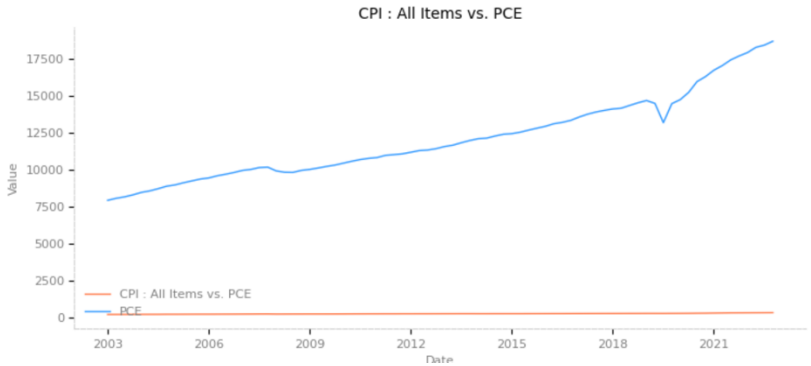
Indicator Measurement Type Harmonization

First we establish a clear understanding of the various types of measurement units present. For instance, our FREDmd Dataset includes a mix of dollar values, counts, rates, ratios and indexes, which will require a different approach for analysis.

```
array(['avg dollars per hour', 'avg hours', 'avg no of weeks',
       'billions of 1982-84 dollars), deflated by cpi',
       'billions of 2012 dollars), deflated by core pce',
       'billions of chained 2012 dollars', 'billions of dollars',
       'billions of dollars, adjusted for inflation and excluding government transfer payments.',
       'exchange rate', 'index = 100',
       'millions of 2012 dollars, deflated by core pce',
       'millions of chained 2012 dollars', 'millions of dollars',
       'thousands of persons', 'percent', 'ratio', 'thousands of units',
       'thousands, seasonally adjusted annual rate'], dtype=object)
```

To do this we load variable metadata from `fredmd_information.csv` obtained for the FRED Definitions document and convert it to a dictionary, which contains information to measurement information for each Indicator. We then map the FRED-MD column indicators to their corresponding measure types.

Hereafter, we standardise certain economic measures by defining conversion factors for different units, e.g., currencies on the same unit scale in billions. This allows us to compare economic indicators reported in other units. However, some indices still are not comparable, showing exponential growth or large fluctuations, as seen below.



Data Transformation with Log and Differencing

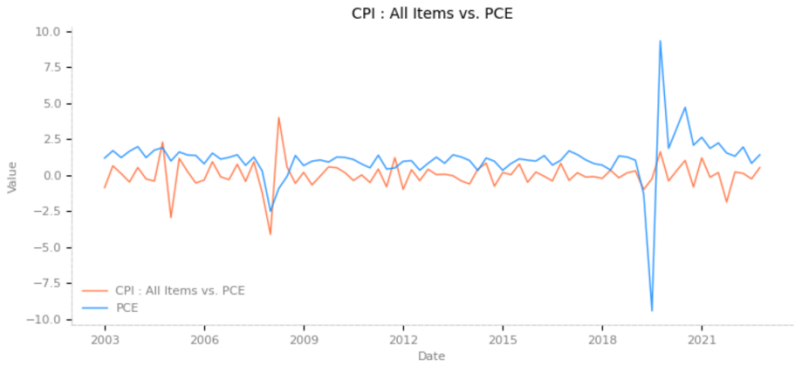
Economic indicators like those often display significant variability over time. In order to stabilise the variance in the dataset for indicators that exhibit exponential growth or large fluctuations. We follow McCracken's suggested transformation types to ensure that our data handling is aligned with 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. *Δxt => x.diff()*
3. **Second Difference:** Captures acceleration or deceleration by examining the change in the first difference. *Δ2xt => 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. *Δ log(xt) => np.log(x).diff()*
6. **Second Difference of Log:** Similar to the second difference but applied to logged data. *Δ2 log(xt) => np.log(x).diff().diff()*
7. **Percentage Change from Prior Period:** Emphasizes relative changes by calculating percentage changes from the previous period. *Δ(xt/xt−1 −1.0) => (indicator / indicator.shift(1) - 1.0) 100*

Implementation:

The process involves mapping the FRED transformation codes to the corresponding series in our `joined_dataset`.

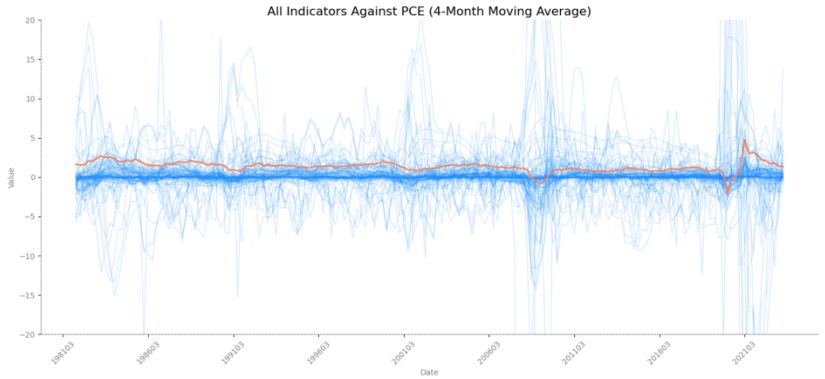
- **Transformation Function:** A specialised function, `modified_log_transform`, applies the selected transformation to each series in the dataset. Each economic indicator is associated with a transformation code that dictates how it should be processed. These codes are retrieved from the `fred_indicator_mappings` dataset.
- **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. Below, you can see the data with the two most correlated indicators with PCE after transformation on a more comparable scale:



By meticulously applying these transformations, we enhance our dataset's suitability for advanced statistical modelling and analysis. This process aligns our methodology with established standards and ensures that each economic indicator is accurately represented, allowing for meaningful comparisons and insights.

Initial descriptive analysis to understand the data:

We initiated our analysis by employing a 4-month moving average to visualise the interplay between PCE and other economic indicators, smoothing out short-term volatilities to reveal long-term trends. This approach highlighted the messy and nuanced influence of economic activities on consumer spending and underscored the prominent role of PCE amidst other indicators.



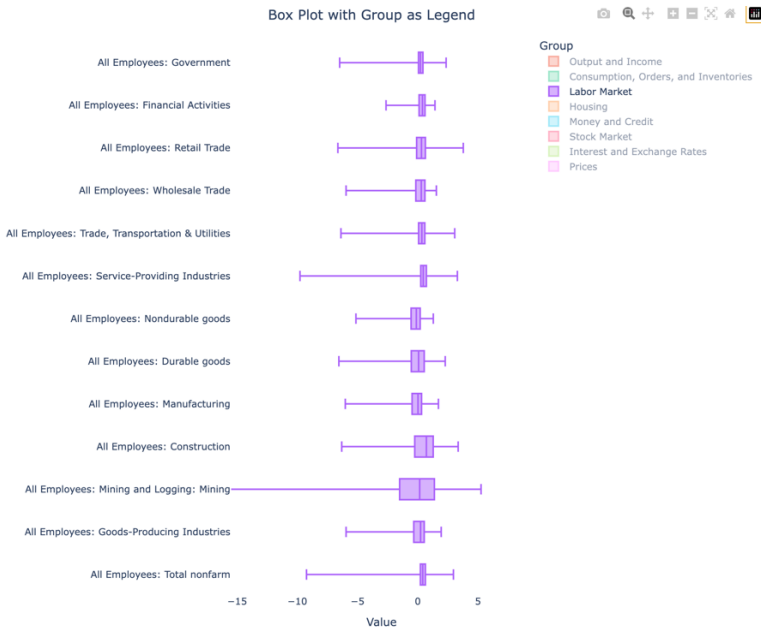
Descriptive Statistics Investigation

The descriptive statistics for economic groups such as Consumption, Orders, Inventories, Housing, and the Labor Market provided insight into their behaviours. For instance, the Consumption, Orders, and Inventories group, with a mean of 0.737 and a high standard deviation of 3.934, exhibited considerable volatility, reflecting diverse impacts on consumer expenditure.

Volatility and Distribution Analysis

Our further dive into volatility analysis, mainly focusing on groups like Money and Credit and the Labor Market, revealed their high susceptibility to rapid economic changes. Such volatility underscores the complex dynamics of how monetary policy effects and employment trends influence consumer spending.

Simultaneously, we mapped indicators to their respective economic groups, enabling a structured analysis conducive to understanding distribution characteristics. An interactive Box plot using plotly was used to visualise and examine variations in median values, spreads, and outliers, providing a granular view of the economic indicators per group.



Key Observations and Economic Implications

- The Labor Market: Civilians Unemployed, Initial Claims and Money and Credit groups emerged as highly volatile, suggesting a strong linkage with consumer confidence and spending behaviours.
- Housing indicators demonstrated stability, hinting at their reliability but perhaps less sensitivity to immediate economic shifts.
- Indicators with high variance, such as Reserves Of Depository Institutions and Crude Oil Prices, were pinpointed as potential early warning signals for changes in consumer spending, albeit warranting caution due to their pronounced volatility.

Visual Temporal Comparison:

To understand how these indicators correlate with PCE over time, we generated line graphs functions with subplots for the top_n indicators against PCE based on their absolute Spearman correlation score. This enabled us to compare their movements over the selected period visually. For this, we leverage a custom function from our `utils.visualisation` module.

Observations from Preliminary Analysis

The initial line graphs juxtaposing top indicators with Personal Consumption Expenditures (PCE) highlights a notable relationships with the labour market.

Integration of Insights for Proxy Validation

Integrating the descriptive statistics and the deeper dives into individual indicator volatilities lays a robust groundwork for understanding the complexities of predicting consumer expenditure. Our findings, from the visualisation of indicator trends to the detailed volatility analysis, highlight the promise of certain groups and indicators as proxies for nowcasting PCE. The varying degrees of volatility and their economic impacts underscore the necessity of a nuanced approach to proxy selection.

Correlation Analysis

We opted to use Spearman's rank correlation, a non-parametric measure that is particularly adept at deciphering the linear relationships without assuming their nature, making it an ideal choice for economic data prone to non-linear trends and outliers. This method's resilience to outliers and its capacity to handle NaN values by omitting them ensures our analysis remains robust and reliable. The Key Objectives of Correlation Analysis is to:

- Identifying Influential Indicators: By sorting correlations from the highest to the lowest based on their absolute values, we pinpoint indicators that exhibit strong linear relationships with PCE.
- Navigating the Correlation Landscape: The sorted correlations, retaining their original signs, offer us a dual lens to view the magnitude and directionality of each relationship. This nuanced approach aids in unravelling how each indicator's fluctuations resonate with shifts in PCE.

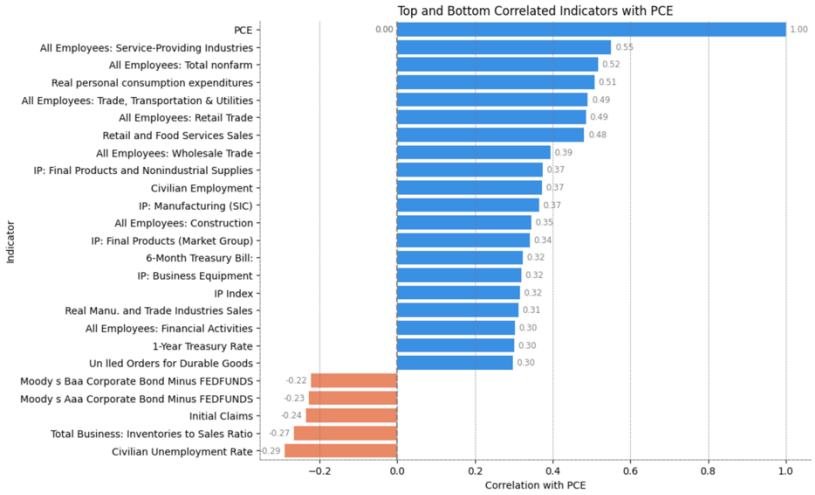
Implementation:

Visualization Strategy: Employing a horizontal bar plot, we delineate positive correlations in sky blue and negative correlations in coral, with a distinct zero correlation marker. This visual distinction underscores the directional influence of each indicator on PCE.

Analytical Precision: By focusing on the top N positively and bottom N negatively correlated indicators, we streamline our investigation towards those variables that present the most substantial predictive value for PCE.

Correlation Analysis Results

- **Labor and Housing Markets:** A significant presence of labour market indicators among the top correlated variables underscores their pivotal role in consumer expenditure dynamics.
- **Moderate to Weak Correlations:** The top correlated indicators exhibit moderate positive relationships with PCE (coefficients ranging from 0.39 to 0.55), while the least correlated indicators show weak negative relationships (coefficients from -0.13 to -0.29).



Multicollinearity Analysis with Variance Inflation Factor (VIF)

Multicollinearity: Assessing the degree to which indicators are interrelated is essential. High multicollinearity among variables can distort the true relationship with PCE, making it difficult to isolate the impact of individual indicators.

We adopt a two-pronged analytical approach to investigate multicollinearity among economic indicators: Circular Correlation Heatmap visualisation and Variance Inflation Factor (VIF) analysis. This comprehensive strategy enables us to identify and address multicollinearity, refining our econometric modelling process.

Circular Correlation Heatmap Visualization

The Circular Correlation Heatmap is a useful tool in our analytical arsenal, offering a holistic view of the interrelationships between economic indicators. This visualisation technique, leveraging the hierarchical clustering of correlation matrices, illuminates the strength and direction of correlations in a visually intuitive manner.

Comprehensive Insights: The heatmap provides an overarching view of indicator correlations, showcasing their interplay and the magnitude of their relationships.

Multicollinearity Detection: It adeptly highlights clusters of tightly correlated variables, allowing for the easy identification of potential multicollinearity among groups of indicators.

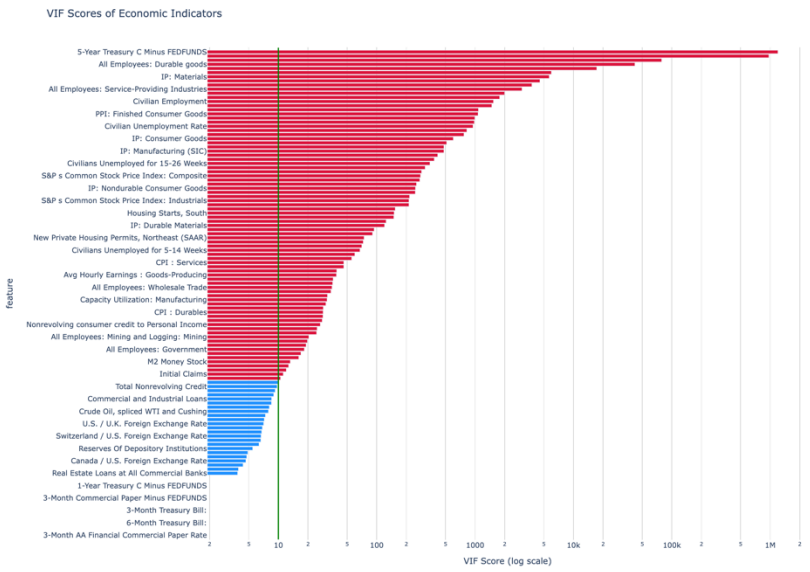
Simplified Interpretation: We turned out focus only to the top applying a threshold that helps distil the heatmap into easily interpretable clusters, emphasising significant relationships while filtering out noise. The threshold determines how many indicators to include from the

Through this visualization, we've observed pronounced collinearity within labour and housing market indicators, indicating potential parallel movements or mutual influences. Such insights are invaluable for pre-empting multicollinearity issues in our models.

Addressing Multicollinearity with Variance Inflation Factor (VIF) Analysis

Next we turn to Variance Inflation Factor (VIF) analysis to assess the impact of multicollinearity. VIF quantifies how much the variance of an estimated regression coefficient is inflated due to predictor intercorrelations. A VIF exceeding 10 typically signals problematic multicollinearity, warranting corrective measures in model specification.

Nowcasting Consumer Expenditure:



Observations: Several indicators have been identified with Variance Inflation Factor (VIF) scores exceeding the threshold, indicative of significant multicollinearity risk. This discovery prompts us to carefully approach the integration of these variables into our econometric model to preserve our analysis's integrity and predictive accuracy.

Strategic Steps Forward

While multicollinearity poses challenges, our strategy is to avoid eliminating these high-collinearity indicators hastily. Doing so could inadvertently strip away valuable insights integral to understanding Personal Consumption Expenditures (PCE).

1. Proxy Selection:

We aim to strategically select proxies demonstrating strong correlations with PCE and contribute unique, indispensable insights into our analysis.

linear regression analysis: We'll examine the predictive strength of each indicator on PCE through linear regression, focusing on the R^2 value to gauge the explanatory power of individual variables.

Seasonality: Investigating seasonality involves identifying and measuring regular, predictable patterns within specific time frames. Seasonal fluctuations can significantly influence economic indicators and, by extension, consumer spending patterns.

Stationarity: Understanding if a time series is stationary is crucial, as it affects the validity of many statistical models. Stationarity implies that the statistical properties of the series do not change over time, which is rarely the case in economic data without transformation or differencing.

2. Dimension Reduction:

Upon selecting our base proxies, we'll apply dimensionality reduction techniques to distil the essence of highly correlated variables into a more manageable set of components. Principal Component Analysis (PCA) is a particularly effective tool, enabling us to consolidate overlapping information into fewer, more potent representative factors. This approach mitigates the impact of multicollinearity and enhances our model's interpretability and efficiency.

By adopting this strategic approach to dealing with multicollinearity, we safeguard against the loss of valuable information while ensuring the reliability and accuracy of our predictive models. These carefully considered steps underscore our commitment to developing a robust analytical framework that captures the nuanced dynamics of consumer expenditure.

Linear Regression Analysis for Private Consumption Expenditure (PCE) Determinants

The crux of this analysis is to harness the R^2 (coefficient of determination) metric, which quantifies the strength of linear relationships between each independent variable and PCE. This metric can identify variables that have the power to explain the variance of PCE.

Methodical Approach to Analysis:

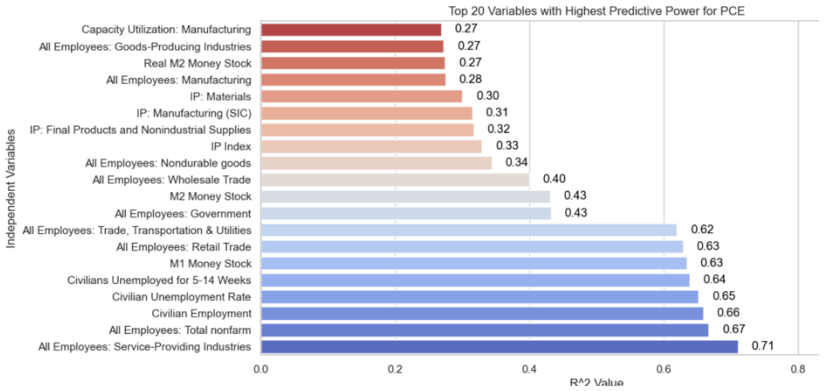
1. Data Preparation:

- Exclusion of Dependent Variable: PCE, being the focus of our study, is set aside from the pool of independent variables to maintain the integrity of our regression model.
- Data Cleansing: This step involves the elimination of rows containing NaN's or infinite values.

Team: Jan Nagtegaal | Aditya Sharma | Evan Trout

GitHub repository: https://github.com/janjannagtegaal/NowCasting-M1

- Modelling: This step involves fitting the linear regression model to each selected variable and PCE, laying the groundwork for precise and insightful analysis.



3. Assessment of R^2 Values:

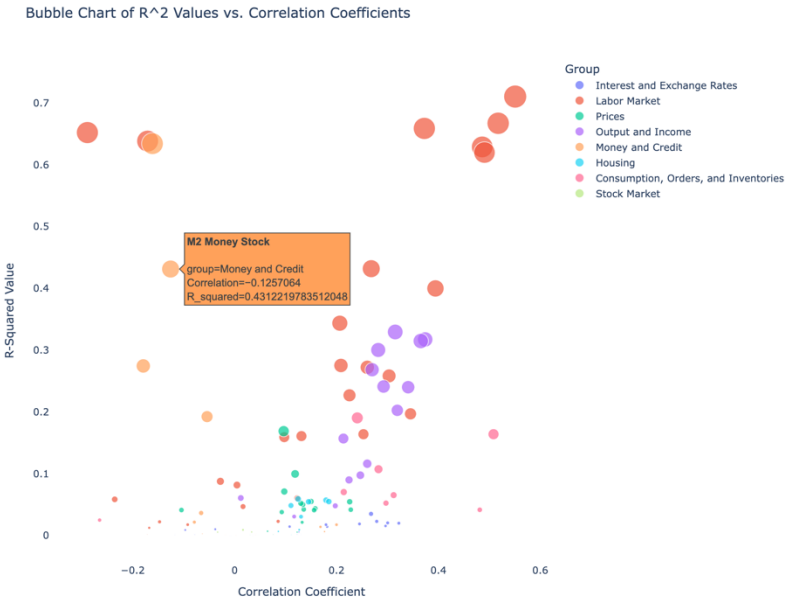
The computation and examination of R^2 values are central to this analysis:

- Prediction and Evaluation: Per model fitting, PCE predictions are generated using each independent variable. The R^2 value for each variable is then calculated, indicating its explanatory power concerning PCE variations.
- Interpretation of Results: A higher R^2 value indicates a stronger linear relationship and a greater extent of variance in PCE explained by the variable. Such variables are earmarked for further analysis, given their potential significance as critical drivers of PCE such as certain labour indicators, M1 & M2 MoneyStock and some IP indexes.

Applying the R^2 metrics enables us to identify the variables most influential in shaping PCE trends. This insight broadens our comprehension of the interplay between indicators and PCE. It assists us in refining our selection process ensuring that our focus is trained on the most impactful determinants of consumer spending.

Utilising R^2 and Correlation Coefficients Together

Interpreting the main findings from our barplot, we can create a interactive scatterplot, which lists each indicator alongside their corresponding R^2 values (coefficient of determination) and correlation coefficients with PCE which can provide a framework for selecting initial proxies for further analysis.



Why This Approach Is Beneficial

1. Identifying Key Drivers of PCE: Indicators with higher R^2 values indicate a stronger linear relationship with PCE, meaning these variables can explain a more significant portion of the variance in PCE. When coupled with the correlation coefficient, which provides direction (positive or negative), we gain a comprehensive understanding of how each variable influences PCE. For instance, "All Employees: Service-Providing Industries" with the highest R^2 value and a strong positive correlation signifies a significant positive influence on PCE.

2. Refining Proxy Selection: The combination of R^2 values and correlation coefficients aids in refining out list of initial proxies.

This dual metric approach facilitates a more nuanced understanding of the relationships between PCE and potential proxies. It allows for identifying the most vital influencers and variables contributing unique insights into PCE dynamics, enriching the overall analysis.

Proxy Selection

Armed with the insights from our R2 and correlation analysis, the next steps are:

Defining Filtering Thresholds: By establishing clear criteria based on correlation coefficients and R2 values, we aim to identify the most informative proxies for our model. We have iteratively selected a threshold of 0.3 threshold for the correlation coefficient and or 0.2 threshold for R2 to obtain the best results.

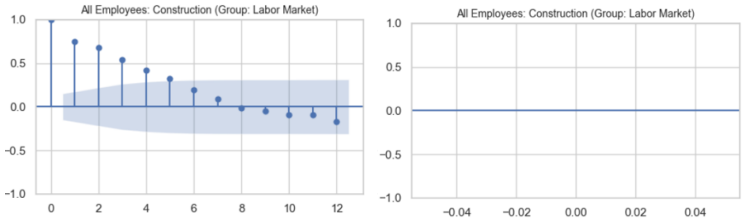
Evaluating Economic Intuition: Beyond statistical measures, we also considered the economic intuition behind each potential proxy. The top indicators are statistically significant and logically connected to PCE dynamics.

Seasonality Assessment

The first critical step after selecting a subset of proxies involves assessing seasonality within our dataset.

- Autocorrelation function (ACF) analysis: We identify indicators with notable seasonality by calculating ACF values for specified lags and identifying those exceeding a predetermined threshold. This step is crucial for recognising patterns that may artificially inflate correlation or regression outcomes.
- Seasonality Removal: Through seasonal decomposition, we adjust our series by stripping away the seasonal component, either additively or multiplicatively. This adjustment yields a series that more accurately represents underlying trends, crucial for subsequent analytical tasks.
- Results and Interpretation: Our initial seasonality check revealed several indicators with significant seasonal patterns. The successful application of seasonal adjustment techniques then mitigated these seasonal influences, as evidenced by the absence of seasonality in the rechecked series. This outcome validates our methodological approach and enhances the reliability of these indicators for further analysis.

Before **After**



Stationarity Assessment

Purpose and Methodology: Ensuring stationarity within our time series data is indispensable for accurate modelling and forecasting. Stationarity implies that the statistical properties of the series do not change over time, a prerequisite for many statistical models. We leverage the Augmented Dickey-Fuller (ADF) test to scrutinise our series for stationarity, focusing on indicators identified as potential proxies for PCE.

- ADF Test: This test assesses whether a unit root is present in the series, with the absence of a unit root (indicated by a p-value below 0.05) confirming stationarity. This step is vital for validating the suitability of our data for further econometric modelling.
- Results and Interpretation: The ADF test outcomes underscored the stationarity of our selected indicators, affirming their appropriateness for in-depth analysis. Notably, the indicators exhibited varying degrees of correlation and R2 values with PCE, enriching our understanding of their dynamics and potential as proxies.

Integration of Findings for Proxy Selection

We've identified a refined list of 31 indicators for modelling PCE by integrating the results from both the seasonality and stationarity assessments with correlation, R^2 and Variance Inflation Factor (VIF) analyses. This selection process prioritises indicators that are not only statistically sound (stationary and devoid of seasonality) but also highly correlated with PCE and explanatory (high R^2 values) while considering multicollinearity (through VIF analysis). Our top 3 proxies identiefied as can be seen below.

Name	Correlation	R_squared	VIF	Test Statistic	P-Value	Conclusion
All Employees: Service-Providing Industries	0.842804	0.710318	2.993945e+03	-7.817825	6.795139e-12	Stationary
All Employees: Total nonfarm	0.816759	0.667095	3.769548e+03	-7.083088	4.612122e-10	Stationary
Civilian Employment	0.811657	0.658787	1.534784e+03	-11.831759	7.960645e-22	Stationary

We realise they have a high level of multicollinearity so to use them with other we will have to use another method to address the collinearity.

Principal Component Analysis (PCA) Analysis: Dimension Reduction

To refine our predictive models, we implemented Principal Component Regression (PCR) analysis, a sophisticated technique that amalgamates Principal Component Analysis (PCA) with Linear Regression. This approach addresses multicollinearity among predictors and harnesses dimensionality reduction to enhance model performance.

Methodological Overview

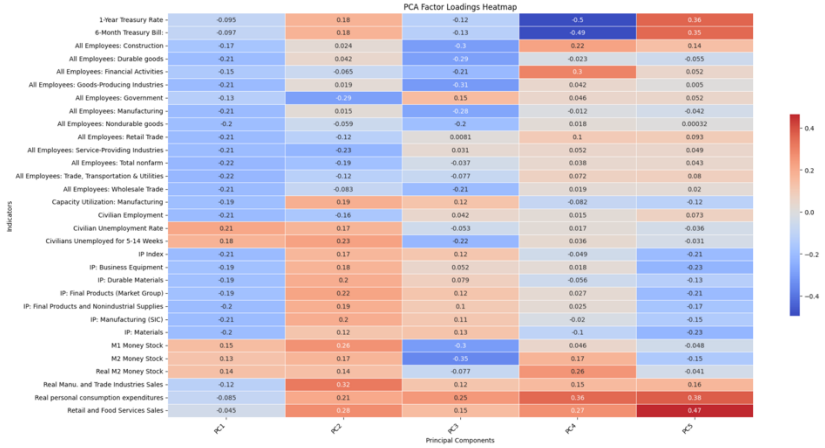
1. Data Preparation: Initial steps focused on the dataset's cleanliness, ensuring that missing values were appropriately handled. This process is vital for maintaining the

accuracy of PCA and subsequent regression analysis, as missing data can significantly distort the results.

2. PCA for Dimensionality Reduction: PCA was performed on the predictors to tackle multicollinearity and reduce our dataset's complexity. This step transformed the original variables into a set of linearly uncorrelated components, known as principal components, which were then used as new predictors.

3. Implementation and Results: The PCR model was operationalised through a pipeline integrating StandardScaler for data normalisation, PCA for dimensionality reduction, and Linear Regression for prediction.

4. Factor Loadings Analysis: Examining the factor loadings revealed how the original variables contributed to each principal component. indicators.



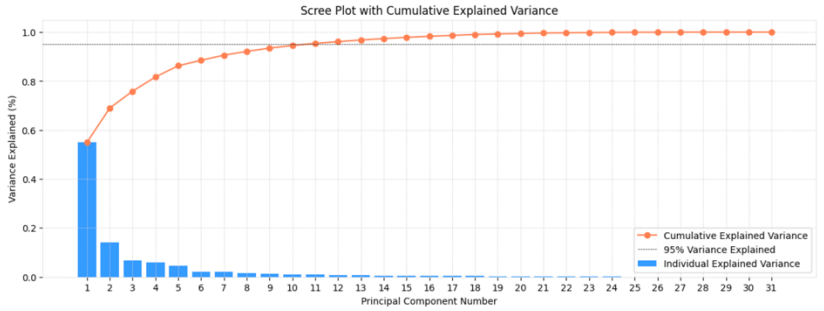
5. Model Performance: The model's predictive accuracy was quantified using:

- Mean Absolute Error (MAE): 0.2832
- Root Mean Squared Error (RMSE): 0.4398

Given the scale of our data, this level of MAE indicates a relatively high accuracy in the model's predictions while the relatively low RMSE further confirms the model's effectiveness in capturing the underlying trends in consumer spending.

5. Scree Plot Analysis: To quantify the contribution of each principal component towards explaining the variance in the dataset, we examined the defined variance ratio. This analysis is encapsulated in the Scree Plot, which visually represents the proportion of the dataset's variance that each principal component accounts for.

The Scree Plot revealed a rapid decline in variance explained by successive principal components, with the initial components capturing the most significant portion of the variance.



3. Regression Using Principal Components: 5 principal components were used as predictors in a Linear Regression model to forecast PCE. This approach leverages the transformed dataset to provide clear, interpretable insights while maintaining the essence of the original data.

We integrated the actual PCE data with the model's predictions to contextualise our model's performance and associated uncertainty. Visualised through a fan chart, this integration illustrates the model's forecasts alongside actual PCE values, with shading indicating the estimated prediction uncertainty.