Nowcasting Consumer Expenditure:

Uncovering Reliable Proxies for Consumer Spending Behavior.

Why This Project Was Chosen

The Challenge

In today's fast-paced world, the delay in economic reporting, such as the quarterly Gross Domestic Product (GDP) figures, poses a significant challenge. Policymakers and market analysts often find themselves making decisions with data that's already out of date. This situation is akin to driving with a rear-view mirror; it tells you where you've been, not where you're going.

The Spark of an Idea

This project was inspired by a straightforward question: ***How can we get current economic data faster?*** Traditional economic indicators serve their purpose but don't keep pace with the rapid changes in consumer behaviour and the broader economy. We aimed to address this gap by leveraging high-frequency data proxies, which offer more immediate insights into consumer spending patterns.

The Approach

Our focus is on harnessing modern data sources to provide a more accurate picture of the economy as it stands today. Throughout the project, we adhered to the maxim: "Define the problem before seeking a solution" and maintained a questioning attitude towards our data. Our journey through problem formulation, data collection and cleaning, analysis, modelling, and finally, presenting results has been guided by a framework that balances technical expertise and a deep understanding of the economic context.

Building a Solution

This initiative required a multidimensional approach, blending expertise from various data science and economics domains. This project is about building a solution that bridges the gap between the need for timely economic data and traditional data collection and reporting methods. This is not about reinventing the wheel but rather about making sure the wheel is rolling in real-time, keeping up with the fast-moving economic landscape.

We're identifying high-frequency data proxies that could paint a clearer, more immediate picture of consumer spending and economic health. Our goal is to ultimately create a tool that provides actionable insights, helping to equip policymakers, businesses, and analysts with the information they need to make informed decisions in real time.

An Invitation to Look Closer

We invite you to explore the potential of real-time economic data with us. This project represents a step forward in making economic data analysis more relevant and timely. It's about understanding the present as clearly as we've come to understand the past, ensuring decision-makers have a clearer view of the economic road ahead.

Specific Questions or Goals

The project was initiated with a deep dive into the economic context, defining the problem statement as needing timely economic data. This foundational step was crucial for setting the direction of our research and analysis. Several key questions aimed at enhancing our understanding of real-time economic dynamics were developed, such as:

- 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, brightness of cities at night, port 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*

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

	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 Pre-processing 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

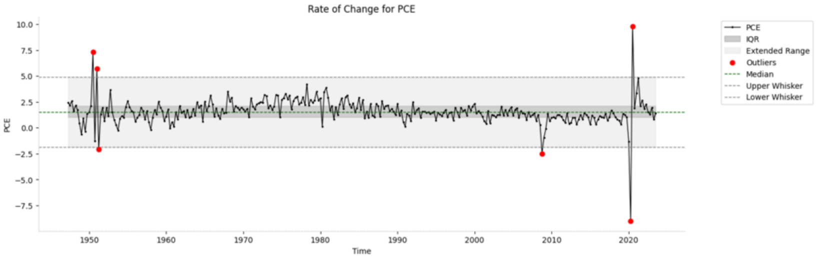
- Hierarchical Naming: Constructs a structured naming system based on indentation levels as the CSV contains indentation to show component hierarchies.
- Abbreviation Mapping: We implement a mapping from full component descriptions to their abbreviations, aiming to simplify and shorten the description 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 format ('YYYYQX') for more accurate joining with the FRED database.
- 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.

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 behaviour, identify any unusual values, and assess its trend over time. We did this by calculating the rate of change of PCE over each period, then calculating the range, IQR, median and outliers that could be visually inspected.



Loading and Pre-processing 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').

Join and Pre-processing the Joined Dataset

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. The quarterly frequency was then transformed to the last month of each quarter in datetime format.

Data Integrity Checks

- Verify Data Consistency with Column Labels: Identify and correct any discrepancies or typographical errors in column names
- Ensure Adherence to Field-Specific Rules: Validate that data within each field complies with specific rules or constraints inherent to its economic significance (e.g., non-negative values for certain indicators).
- Duplicate Entries: Identify and remove any duplicate rows within the dataset to ensure the uniqueness of each data entry.
- Ensure Proper Data Types for Each Column: Confirm that each column is assigned the correct data type, we converted indicator columns with integers types to floats and ensuring datetime formats are correctly applied for indices.
- Alignment Between Dataset Columns and Definitions Dataframe: Ensure a one-to-one correspondence between columns in the main dataset and the definitions or mappings dataframe, which contains essential metadata like transformation codes and economic groups.

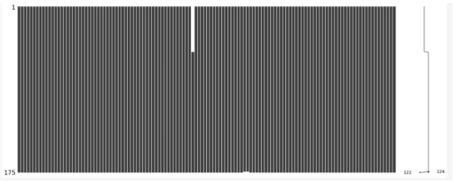
Data Filtering

Date Range Filtering: Considering the Law of Large Numbers (LLN) for stability in our analysis, we selected data from 1980 onwards to ensure a broad representation of business cycles, enabling us to aggregate extensive data from various sources, smoothing out individual data point noise. This approach enhances our dataset's ability to accurately reflect the economic indicators' true state, acknowledging the unique traits of economic data such as cyclical patterns, trends, and volatility.

Column filtering: We also removed columns deemed less reliable, as highlighted by McCracken in their 2015 working paper for the Federal Reserve Bank of St. Louis, such as the "Help-Wanted Index for the United States" and the "Ratio of Help Wanted/No. Unemployed" due to their reliability concerns. Additionally, we dropped 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 suffers from a one-year lag in recent data availability in the FRED database.

Handling Missing Data

We used the missingno to visualise the missing values in 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

We applied the Z-score method to identify and replace outliers with np.nan values for each indicator column. Data points with a Z-score exceeding a threshold of 3 were considered outliers.

COVID-19 outliers: To identify underlying economic trends, it could be helpful to exclude the COVID-19 pandemic period (2020-2021). This would eliminate the distortions caused by this event and maintain historical consistency. By doing this, we can ensure that the atypical anomalies of the pandemic do not skew models, improving the accuracy of long-term forecasting.

For our economic analysis, however, we decided to include the COVID-19 pandemic period (2020-2021) to better understand its impact on economic indicators. This approach enhances model robustness by acknowledging the influence of unprecedented events.

Normalisation

When visually examining the plotted data, significant scale disparities between indicators and PCE become evident, alongside the challenge posed by Normality Bias. It's a common trait of economic datasets to deviate from a normal distribution, requiring meticulous consideration during normalization processes. This bias holds particular relevance in the context of economic time-series data, which is often marked by trends, cycles, and volatility. To effectively mitigate these issues, a thoughtful and precise approach to data transformation is essential.

Indicator Measurement Type Harmonization

First, we inspect the various measurement units present by loading 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 better compare economic indicators reported in different units.

Data Transformation with Log and Differencing

In our endeavour to harmonize the measurement types across the broad spectrum of economic indices, some still are not comparable, showing exponential growth or large fluctuations, as seen below.

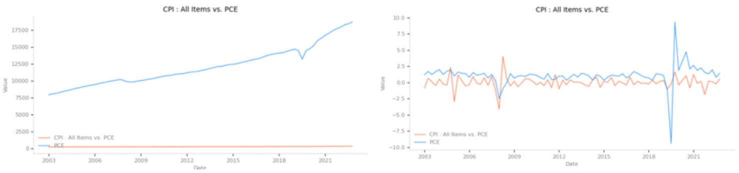
We follow McCracken's suggested transformation types to ensure that our data handling is aligned with established economic analysis practices. This promotes accuracy and consistency in our analyses. Transformation Types as per FRED column tcode denotes the following data transformation for a series x:

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*

As we map the FRED transformation codes and apply various transformations to stabilise variance and linearise growth trends, we remain cognizant of the limitations imposed by Normality Bias. 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.

Resulting Adjustments: The transformed data is then processed. Below, you can see the transformed data before and after transformation on a more comparable scale:

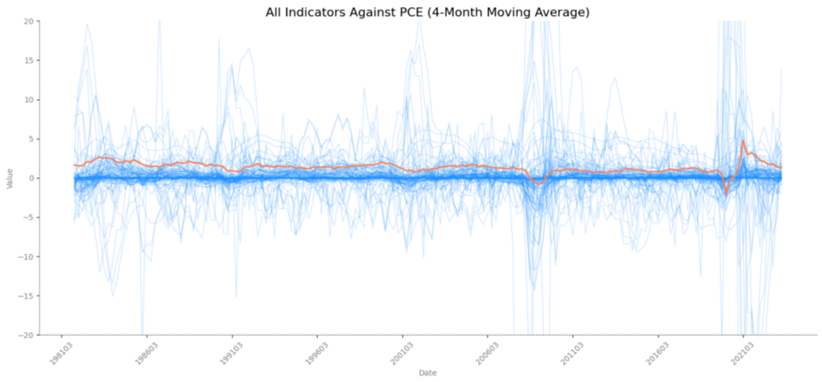


By meticulously applying these transformations, we enhance our dataset's suitability for advanced statistical modelling and analysis. This process aligns our methodology with established standards and ensures that each economic indicator is accurately represented, allowing for meaningful comparisons 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 noisy and nuanced influence of economic activities on consumer spending and underscored the prominent role of PCE amidst other indicators.

Nowcasting Consumer Expenditure:



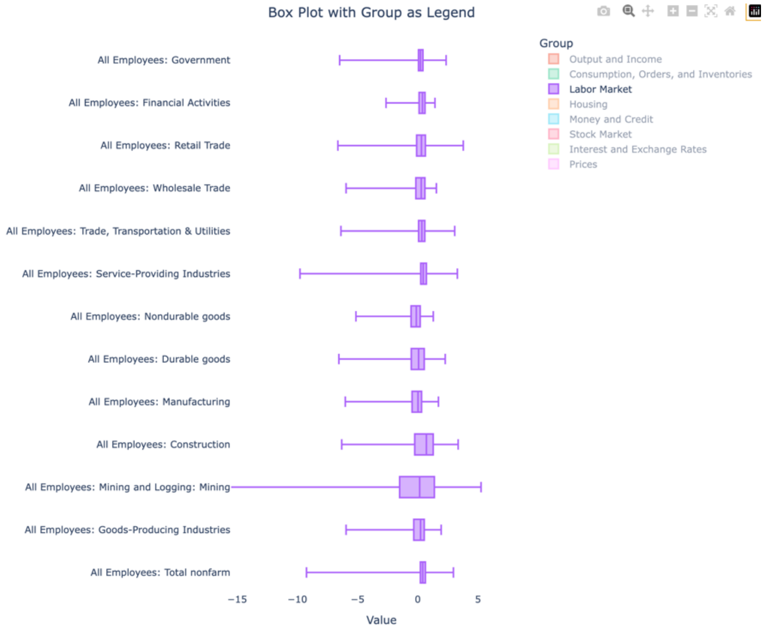
Descriptive Statistics Investigation

The descriptive statistics for economic groups such as Consumption, Orders, Inventories, Housing, and the Labour 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.



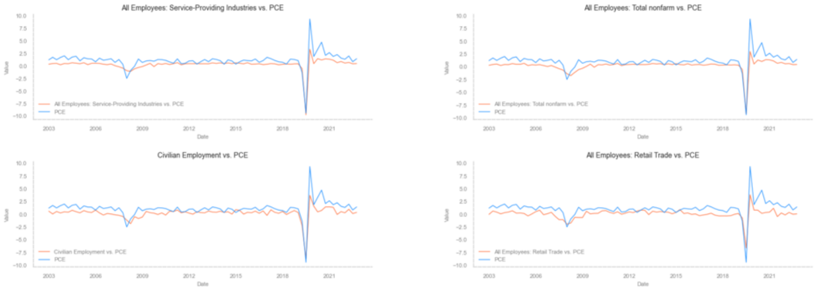
Key Observations and Economic Implications

- Labour 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.
- Reserves Of Depository Institutions and Crude Oil Prices showed high variance and, as such 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 behave with PCE over time, we generated line graphs functions with subplots for the top_n indicators against PCE based on their absolute 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.

Nowcasting Consumer Expenditure:



Observations from Preliminary Analysis

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

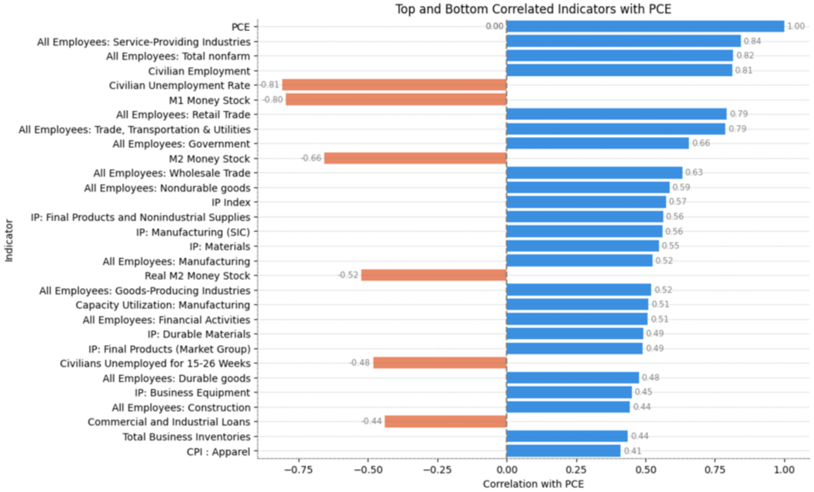
By integrating the descriptive statistics with deeper dives into individual indicator volatilities, we can lay a strong foundation to comprehend the complexities of predicting consumer expenditure. Our visualisation of indicator trends and detailed volatility analysis reveal the potential of certain groups and indicators as proxies for nowcasting PCE. The varying degrees of volatility and their economic impacts emphasise the need for a meticulous approach to proxy selection. Our findings underscore the significance of this approach in accurately predicting consumer expenditure.

Correlation Analysis

Correlation vs. Causation: A significant risk in economic data analysis is inferring causation from correlation without adequate evidence or control for confounding factors. Assuming a direct cause-effect relationship without considering the complexity of economic systems could lead to biased interpretations.

Implementation:

- 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.
- 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.



Key Findings:

- Strong Positive Correlations: Service-providing industries, Total Nonfarm Employment, and Civilian Employment are highly correlated with PCE, showing coefficients above 0.8. This suggests that employment levels, particularly in service sectors, are closely tied to consumer spending.
- Significant Negative Correlations: The Civilian Unemployment Rate and M1 Money Stock show strong inverse relationships with PCE, with coefficients around -0.8 and -0.79 respectively. Higher unemployment rates and fluctuations in the money stock inversely affect consumer spending.
- Moderate to Mild Correlations: Indicators such as Retail Trade, Trade/Transportation/Utilities, and Government Employment have moderate positive correlations, ranging from 0.78 to 0.65, indicating that these sectors also contribute to consumer spending trends but to a lesser extent.
- Industrial Production and Money Stock: Industrial Production indices and Real M2 Money Stock exhibit correlations around 0.57 to -0.52 with PCE. These suggest that manufacturing output and broader monetary conditions have a measurable impact on consumer expenditures.
- Lower Correlations: Sectors like Manufacturing, Durable Goods, and Construction show lower positive correlations (0.52 to 0.44), pointing to a less direct but still significant relationship with PCE.

Team: Jan Nagtegaal | Aditya Sharma | Evan Trout

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

Implications:

- Labor Market Dynamics: The strong correlation between PCE and employment indicators underscores the critical role of labor market health in driving consumer spending, emphasizing the importance of job creation and stability in service and nonfarm sectors.
- Economic Policy Considerations: The negative correlations with unemployment and money stock highlight the sensitivity of consumer spending to monetary supply and labor market conditions, suggesting that policies aiming to reduce unemployment and stabilize the money supply could positively influence PCE.
- Sector-Specific Insights: Moderate correlations in sectors like retail and government employment indicate that while these areas contribute to consumer spending, their impact may be influenced by broader economic conditions or specific sectoral trends.

This analysis illustrates the interconnectedness of various economic indicators with consumer spending, offering valuable insights for policymakers, analysts, and investors aiming to understand or forecast PCE trends.

Multicollinearity Analysis with Variance Inflation Factor (VIF)

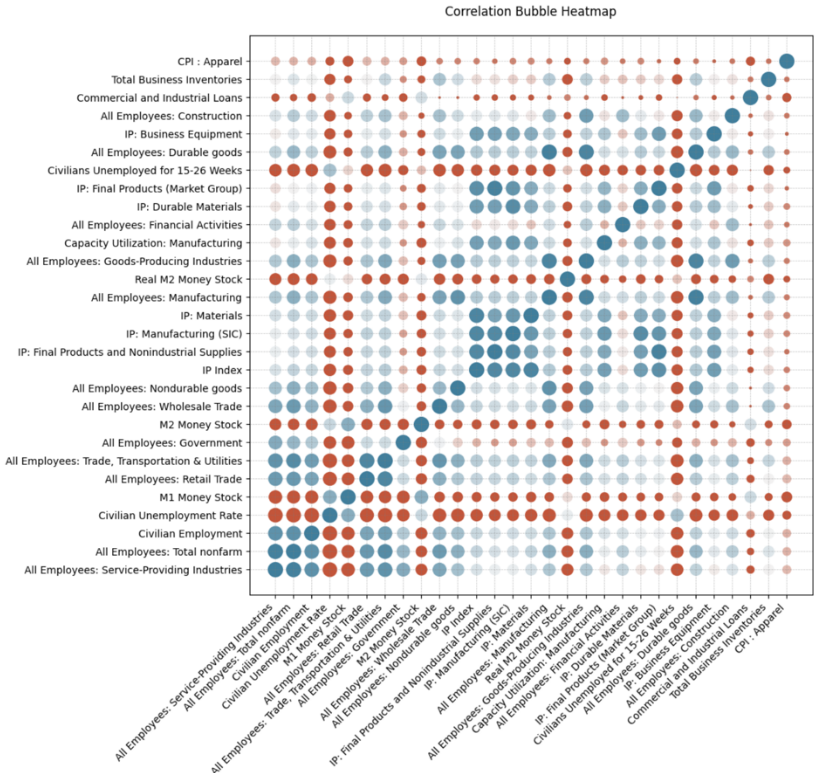
Multicollinearity: Assessing the degree to which indicators are interrelated is essential. High multicollinearity among variables can distort the true relationship with PCE, making it difficult to isolate the impact of individual indicators.

Ignoring Multicollinearity: Multicollinearity among predictors can lead to overfitting. It indicates that one or more independent variables in a regression model can be linearly predicted from the others with a substantial degree of accuracy. This can inflate the variance of the coefficient estimates and make the model more sensitive to changes in the model's specifics.

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 provides an overarching view of indicator correlations, showcasing their interplay and the magnitude of their relationships. It adeptly highlights clusters of tightly correlated variables, allowing for the easy identification of potential multicollinearity among groups of indicators.

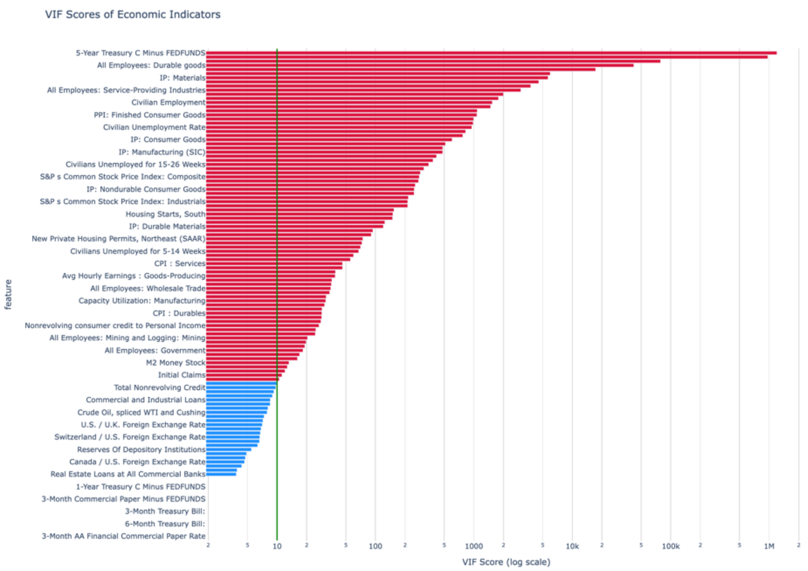


Through this visualization, we've observed pronounced collinearity within labour and other market indicators, indicating potential parallel movements or mutual influences. Such insights are invaluable for pre-empting multicollinearity issues in our models.

Determining 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 with Variance Inflation Factor (VIF) scores exceeding the threshold indicate 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 approach is to avoid eliminating these high-collinearity indicators hastily. Doing so could inadvertently strip away valuable insights integral to understanding Personal Consumption Expenditures (PCE). 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 as seasonal fluctuations can significantly influence economic indicators.

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.

Proxy Selection: We will define a threshold for filtering proxies by their correlation and R^2 power over PCE on the remaining Proxies.

Conduct further hypothesis testing on the linear relationship between indicators and PCE and the distribution of the residuals.

Dimension Reduction: we'll apply Principal Component Analysis (PCA) for dimensionality reduction on our final proxies to enable us to consolidate overlapping information into fewer, more potent representative factors.

Linear Regression Analysis

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.

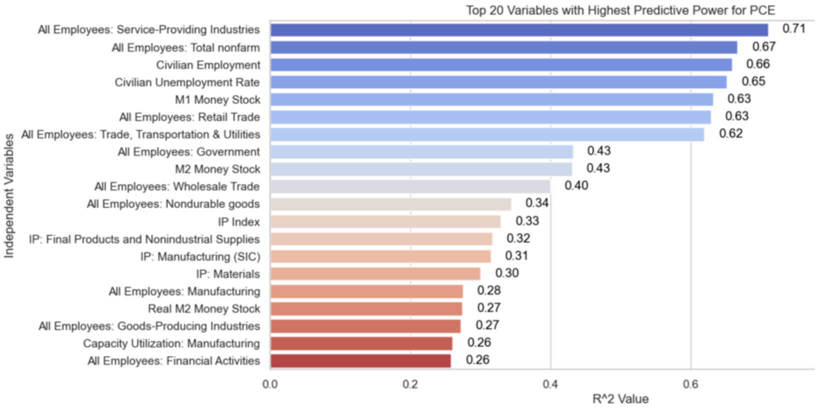
Methodical Approach to Analysis:

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

Team: Jan Nagtegaal | Aditya Sharma | Evan Trout

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



2. Assessment of R^2 Values:

- **Substantial Influence:** Indicators such as Service-Providing Industries, Total Nonfarm Employees, and Civilian Employment, with R^2 values greater than 0.65, highlight a substantial influence on PCE. The strong positive correlations further affirm the pivotal role of employment sectors in driving consumer spending.
- **Influence of Unemployment and Money Stock:** The Civilian Unemployment Rate and M1 Money Stock, with negative correlations and R^2 values indicating moderate to high explanatory power, illustrate how changes in these areas can inversely impact PCE. These relationships underscore the sensitivity of consumer spending to broader economic conditions.
- **Sector Diversity:** The broad range of sectors correlated with PCE, reflected in both positive and negative correlations across retail, manufacturing, and government employment, showcases the multifaceted drivers of consumer spending. This diversity points to the complex interplay between different economic activities and PCE.

3. Why This Approach Is Beneficial

Identifying Key Drivers of PCE: Indicators with higher R^2 values indicate a stronger linear relationship with PCE and 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.

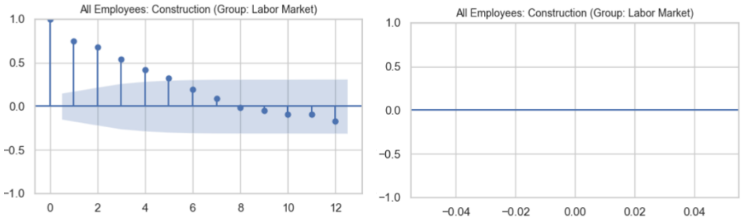
This dual metric approach facilitates a more intricate understanding of the relationships between PCE and potential proxies. It allows for identifying the most vital influencers and variables contributing unique insights into PCE dynamics, enriching the overall analysis.

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.
- **Results and Interpretation:** Our initial seasonality check revealed several indicators with significant seasonal patterns. The successful application of seasonal adjustment techniques mitigated these seasonal influences, as evidenced by the absence of seasonality in the rechecked series.

Before After



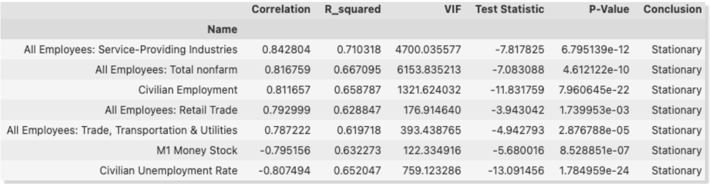
Stationarity Assessment

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. The challenge is compounded by the Normality Bias inherent in economic datasets. These datasets frequently diverge from the normal distribution, exhibiting trends, cyclicality, and volatility that reflect the complex dynamics of economic indicators over time. 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 R^2 values with PCE, enriching our understanding of their dynamics and potential as proxies.

Integration of Findings for Proxy Selection

We've filtered the dataset for proxies to have a correlation coefficient of 0.6 and R² of 0.5. from this we have identified a refined list of 7 indicators for modelling PCE.



	Correlation	R_squared	VIF	Test Statistic	P-Value	Conclusion
Name						
All Employees: Service-Providing Industries	0.842804	0.710318	4700.035577	-7.817825	6.795139e-12	Stationary
All Employees: Total nonfarm	0.816759	0.667095	6153.835213	-7.083088	4.612122e-10	Stationary
Civilian Employment	0.811657	0.658787	1321.624032	-11.831759	7.960645e-22	Stationary
All Employees: Retail Trade	0.792999	0.628847	176.914640	-3.943042	1.739953e-03	Stationary
All Employees: Trade, Transportation & Utilities	0.787222	0.619718	393.438765	-4.942793	2.876788e-05	Stationary
M1 Money Stock	-0.795156	0.632273	122.334916	-5.680016	8.528851e-07	Stationary
Civilian Unemployment Rate	-0.807494	0.652047	759.123286	-13.091456	1.784959e-24	Stationary

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 R² values) while considering multicollinearity (through VIF analysis).

Hypothesis Testing for Economic Indicators

Next, we will focus on the relationship between each final proxy and PCE to determine if these features significantly predict PCE movements. Two primary aspects will be tested:

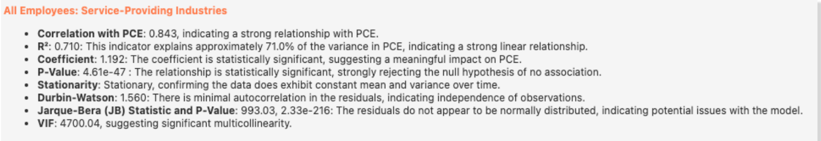
Linear Relationship with PCE: the existence of a significant linear relationship with PCE:

- Null Hypothesis: There is no significant linear relationship between the Proxy and PCE.
- Alternative Hypothesis: The Proxy and PCE have a significant linear relationship.

Distribution of Residuals: The normality of residuals is tested to validate the linear regression model's assumptions.

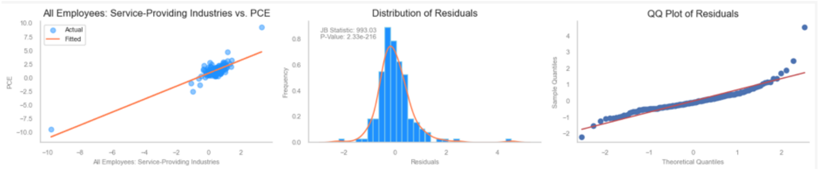
- Null Hypothesis: The residuals from the regression model are normally distributed, indicating the model's assumptions about the error term distribution are valid.
- Alternative Hypothesis: The residuals from the regression model are not normally distributed, suggesting potential issues with the model such as misspecification or the presence of outliers.

For each indicator, we generated a fitted linear regression scatter plot, a distribution of the residuals plot and a QQ plot of the residuals with an interpretation of the results, which can be accessed in the final notebook.



All Employees: Service-Providing Industries

- **Correlation with PCE**: 0.843, indicating a strong relationship with PCE.
- **R²**: 0.710: This indicator explains approximately 71.0% of the variance in PCE, indicating a strong linear relationship.
- **Coefficient**: 1.192: The coefficient is statistically significant, suggesting a meaningful impact on PCE.
- **P-Value**: 4.61e-47 : The relationship is statistically significant, strongly rejecting the null hypothesis of no association.
- **Stationarity**: Stationary, confirming the data does exhibit constant mean and variance over time.
- **Durbin-Watson**: 1.560: There is minimal autocorrelation in the residuals, indicating independence of observations.
- **Jarque-Bera (JB) Statistic and P-Value**: 993.03, 2.33e-216: The residuals do not appear to be normally distributed, indicating potential issues with the model.
- **VIF**: 4700.04, suggesting significant multicollinearity.



Conclusion from a Statistical Perspective: The chosen proxies—ranging from employment sectors to the M1 Money Stock and the Civilian Unemployment Rate—demonstrate strong statistical significance and correlation with Personal Consumption Expenditures (PCE). Notably:

- Correlation Coefficient: Each proxy exhibits a correlation coefficient above 0.78, suggesting a potent linear relationship with PCE.
- R² values: These variables explain 62.0% to 71.0% of the variance in PCE.
- Statistical Significance: the statistical significance of these relationships is further validated by p-values well below the 0.05 threshold, strongly rejecting the null hypothesis of no association.
- Multicollinearity: the analysis also uncovers multicollinearity issues, as evidenced by very high Variance Inflation Factor (VIF) scores and potential model assumptions violations, highlighted by the Jarque-Bera test results.
- Autocorrelation: some instances of autocorrelation as indicated by the Durbin-Watson statistic.

From an economic standpoint, the strong relationship between these labour market indicators and PCE aligns with theoretical expectations.

- Positive correlation: Employment levels and consumer spending are closely linked, as higher employment typically leads to increased disposable income and, consequently, higher consumer spending.
- Negative correlation: The M1 Money Stock and Civilian Unemployment Rates' negative correlation with PCE further corroborates economic theories that relate the money supply and unemployment rates inversely to consumer spending.

Next Steps:

Given the significant multicollinearity among the selected proxies, a logical next step involves employing Principal Component Analysis (PCA) to address this issue. PCA can reduce the dimensionality of the dataset while retaining the variance present in the data, thereby mitigating multicollinearity without substantially losing information.

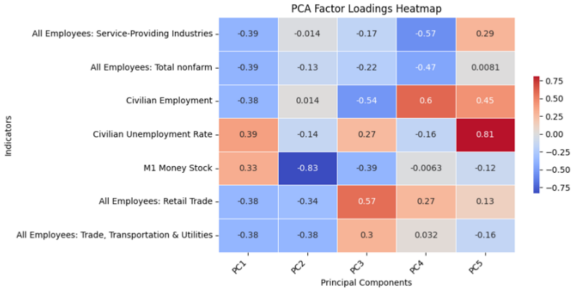
Further, the issues with normality and autocorrelation of residuals necessitate additional model diagnostics and possibly the exploration of alternative modelling approaches or transformations to better meet the assumptions of linear regression. This might include employing generalised linear models, transforming variables, or applying different error correction models to account for autocorrelation.

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 technique will streamline the dataset into principal components that can be used as new, uncorrelated predictors, potentially enhancing model performance and interpretability.

Methodological Overview

- Data Preparation: Initial steps focused on the dataset's cleanliness, ensuring that missing values were appropriately handled.
- 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.
- 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.
- Factor Loadings Analysis: Examining the factor loadings revealed how the original variables contributed to each principal component. indicators.

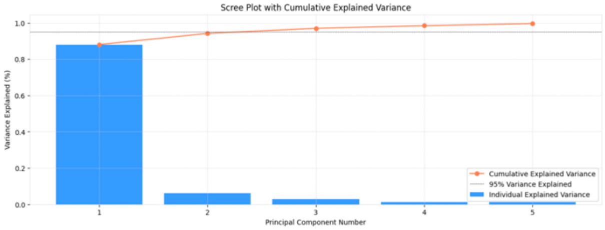


- Model Performance: The model's predictive accuracy was quantified using:
  - Mean Absolute Error (MAE): 0.3132
  - Root Mean Squared Error (RMSE): 0.4069
  - Cross-Validation Performance: (Average MSE): 0.4902

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

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.

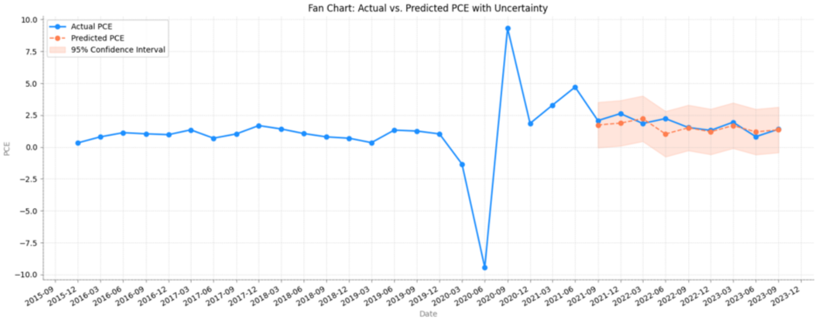


Having too many components can lead to overfitting. Choosing the number of components judiciously, based on the explained variance ratio, is crucial to capture most of the information without overly complicating the model.

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.

Regression Using Principal Components:

2 principal components were used as predictors in a Linear Regression model to forecast PCE. We integrated the actual PCE data with the model's predictions to contextualise our model's performance and associated uncertainty. Visualised through a line chart, this integration illustrates the model's forecasts alongside actual PCE values, with shading indicating the estimated prediction uncertainty.


Fan Chart: Actual vs. Predicted PCE with Uncertainty
Actual PCE
Predicted PCE
95% Confidence Interval
PCE
Date