Nowcasting Consumer Expenditure:

Uncovering Reliable Proxies for Consumer Spending Behaviour.

Why This Project Was Chosen

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. This delay of conventional economic indicators is akin to driving with a rear-view mirror; it tells you where you've been, not where you're going, making it hard to make informed decisions based on current data.

Our approach involves identifying high-frequency data proxies that can provide real-time insights into the economy for policymakers, market analysts, financial institutions, and business leaders based on a rigorous problem-solving methodology.

Objectives

We started the project by exploring the economic context and identifying the need for timely economic data. We developed key questions and goals focused on real-time economic dynamics to guide our research and analysis.

- Identification of high-frequency data sources as accurate proxies for consumer spending.
- Validation of these proxies against established consumer expenditure measures.
- We are 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

Alternative data sources like credit card transactions, retail foot traffic, city brightness at night, port traffic, and online search trends have been explored for economic forecasting. These studies highlight the potential of high-frequency data to predict economic trends in near real-time. We found inspiration from 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

	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.

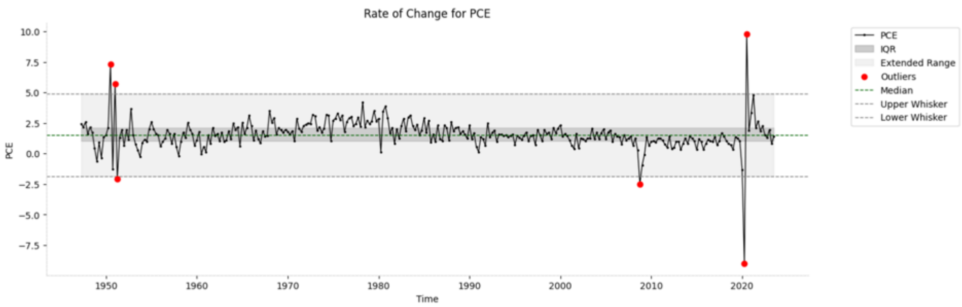
Team: Jan Nagtegaal | Aditya Sharma | Evan Troutt GitHub repository: https://github.com/janjannagtegaal/NowCasting-M1

Nowcasting Consumer Expenditure:

Loading and Pre-processing GDP Data

We loaded GDP data from a CSV file, renamed columns and created a structured naming system, mapped full descriptions to abbreviations, converted date columns to a more standard format, transposed the dataset, and ensured all data columns were numeric for statistical analysis.

Initial Visualisation of PCE. 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 and then calculating the range, IQR, median and outliers that could be visually inspected.



Loading and Pre-processing FRED-MD Data:

Then we retrieved the latest version of the FRED-MD dataset, dropped rows with NAs, converted the 'SASdate' column to a Period Index, mapped FRED-MD column names to their descriptions using a separate definitions file, filtered monthly data to select only the last month of each quarter, and transforms the data into a quarterly format for frequency alignment with our PCE dataset. We also visually inspected a few key indicators using the abovementioned approach.

Integrating and Pre-processing the Joined Dataset

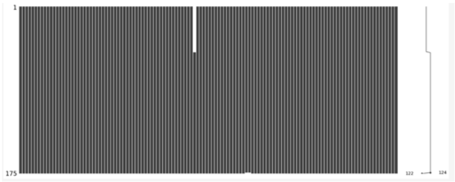
Joining Data Sources: After pre-processing both datasets and aligning temporal frequencies, the FRED-MD dataset was merged with the Personal Consumption Expenditures (PCE) data on their quarterly indices and assigned to joined_dataset. The quarterly frequency was then transformed to the last month of each quarter in datetime format.

Data Integrity Checks: We have completed several integrity checks as part of the data-cleaning process. We verified the column labels and corrected any inconsistencies or errors found. Additionally, we applied field-specific rules to validate data compliance with specific constraints, such as non-negative values for certain indicators. To ensure data uniqueness, we identified and removed any duplicate entries. We also confirmed that each column had the correct data type and applied datetime formats as needed. Finally, to align the dataset columns with definitions, we ensured a one-to-one correspondence between the dataset columns and metadata in the definitions or mappings dataframe, including transformation codes and economic groups.

Data 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' actual state, acknowledging the unique traits of economic data such as cyclical patterns, trends, and volatility.

Column filtering: Due to their reliability concerns, we removed columns deemed less reliable, as highlighted by McCracken in their 2015 working paper for the Federal Reserve Bank of St. Louis.

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: Excluding the COVID-19 pandemic period can help identify long-term economic trends. However, including it in the analysis can provide a better understanding of its impact on economic indicators and enhance model robustness. We opted to include it.

Normalisation

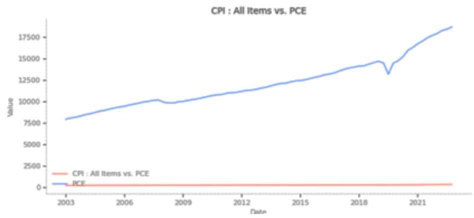
The data shows significant differences in scale between indicators and PCE, and Normality Bias poses a challenge. Economic datasets often deviate from a normal distribution, so normalisation requires careful consideration. This bias is particularly relevant in economic time-series data, marked by trends, cycles, and volatility. A thoughtful and precise approach to data transformation is essential to mitigate these issues.

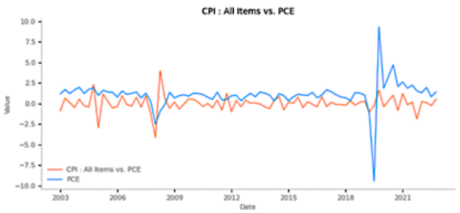
Indicator Measurement Type Harmonization: We gather measurement information for each FRED-MD column indicator and standardise specific economic measures by defining conversion factors for different units. This allows us to compare economic indicators reported in other units.

Data Transformation with Log and Differencing: We use McCracken's recommended transformation types to harmonise measurement types for economic indices. This ensures accuracy and consistency in our analyses, even for series that show large fluctuations or exponential growth. Each economic indicator is associated with a transformation code that dictates how it should be processed as follows:

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
3. Second Difference:	Captures acceleration or deceleration by examining the change in the first difference.	$\Delta 2xt$
4. Natural Log:	Stabilizes variance and linearises exponential growth trends.	$\log(xt)$
5. First Difference of Log:	Transforms data into a stationary series, indicating percentage changes.	$\Delta \log(xt)$
6. Second Difference of Log:	Similar to the second difference but applied to logged data.	$\Delta 2 \log(xt)$
7. Percentage Change from Prior Period	Emphasizes relative changes by calculating percentage changes from the previous period	$\Delta(xt/xt{-}1 - 1.0)$

Implementation: We transformed our dataset and successfully stabilised variance and linearised growth trends while remaining aware of Normality Bias. Below, you can see the transformed data before and after transformation on a more comparable scale:

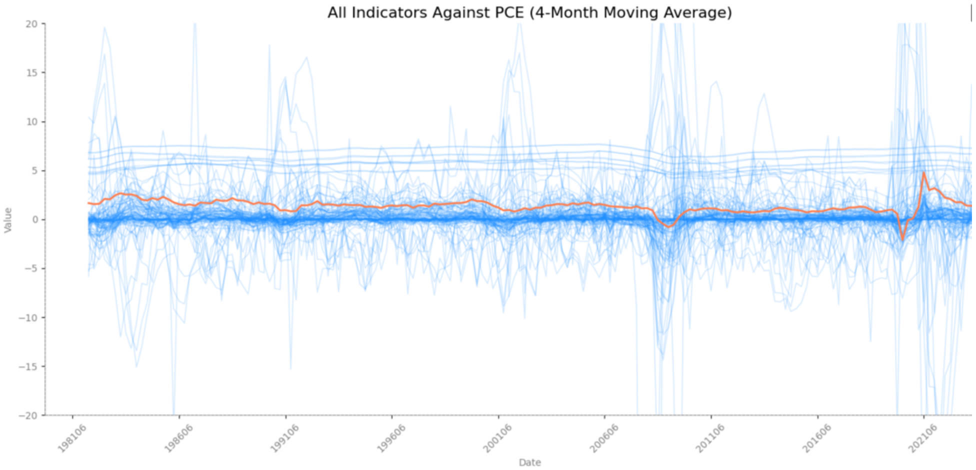




Nowcasting Consumer Expenditure:

Finding clarity from complexity

In the pursuit of distilling clarity from complexity, our initial visualisation below confronts us with a daunting challenge: a dense mosaic of 123 economic indicators, each represented in a subtle shade of faded dodger blue, against the vivid contrast of our target variable, Private Consumption Expenditure (PCE), in bright coral. While initially overwhelming, this visual symbolises the intricate web of economic activity we aim to decipher. Our objective is to meticulously unearth pivotal indicators in this extensive data noise, that hold the key to anticipating the next shifts in PCE. This visual representation serves as a stark reminder of the project's essence.

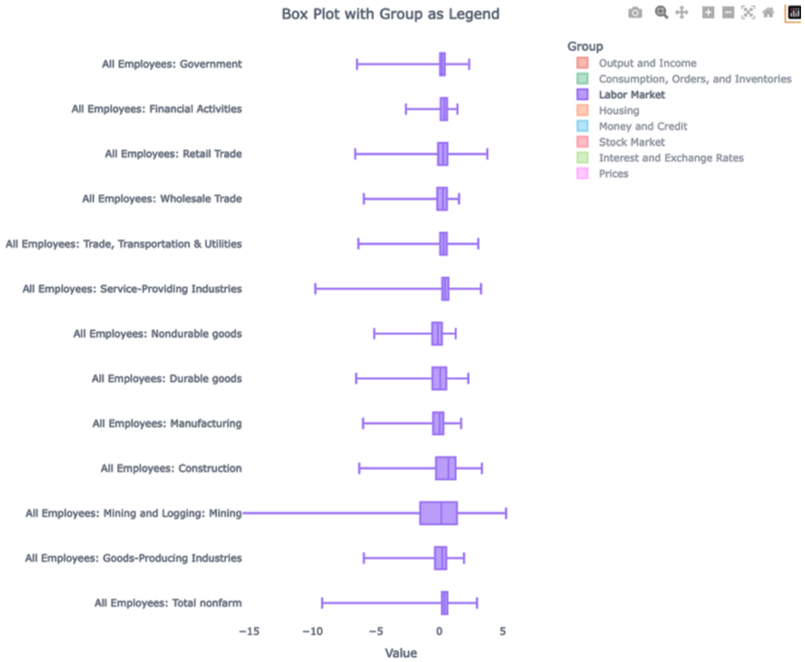


Initial descriptive analysis to understand the data:

Descriptive Statistics Investigation: We performed descriptive statistics for economic groups like Consumption, Orders, Inventories, Housing, and the Labour Market, providing insights into their behaviours. 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. We also performed the same on a more granular level to gauge outlying proxies.

Volatility and Distribution Analysis: Our dive into volatility analysis revealed the high susceptibility of groups like Money and Credit and the Labour Market to rapid economic changes. This underlines the complex dynamics of how monetary policy effects and employment trends influence consumer spending.

Interactive Distribution Visualisation: We mapped indicators to their respective economic groups and visualised them using an interactive Box plot via Plotly. This provided a granular view of the economic indicators per group, including variations in median values, spreads, and outliers.



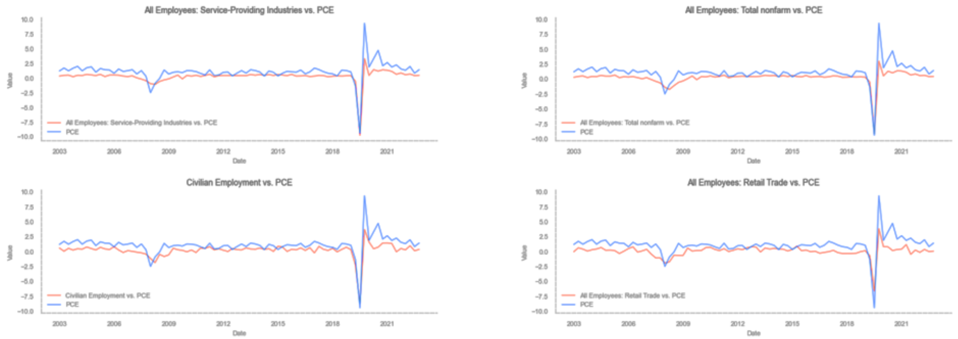
Key Observations and Economic Implications

- 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.

Nowcasting Consumer Expenditure:

- 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. They were pinpointed as potential early warning signals for changes in consumer spending, albeit warranting caution due to their pronounced volatility.

Visual Temporal Comparison: To compare how top_n indicators move with PCE over time, we generated line charts with subplots based on their absolute correlation score. This allowed us to analyse their movements over the selected period visually.



Observations from Preliminary Analysis: Our analysis of top economic indicators and Personal Consumption Expenditures (PCE) revealed strong relationships with the labour market. However, carefully selecting proxies is crucial due to their varying volatility levels and economic impacts.

Correlation Analysis

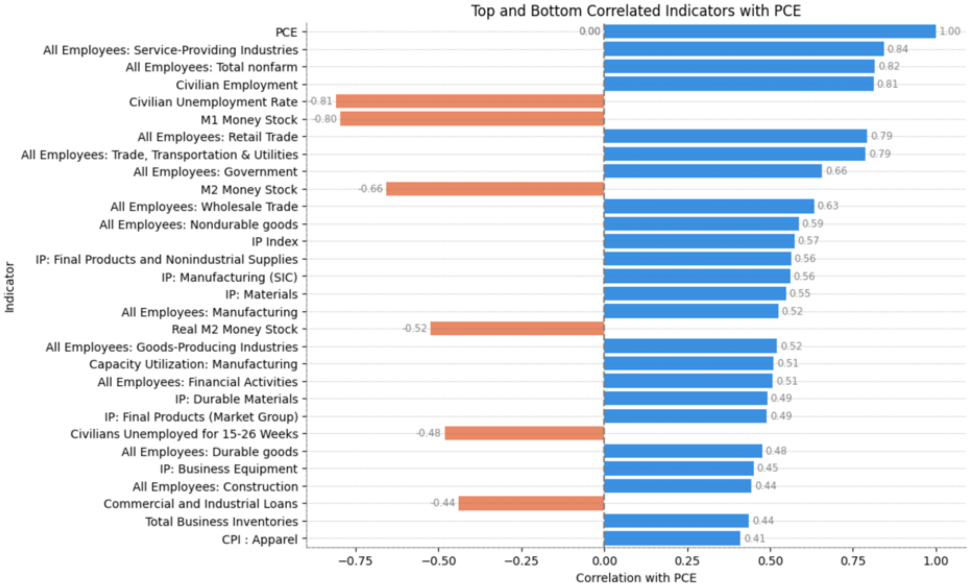
Correlation vs. Causation: A significant risk in economic data analysis is inferring causation from correlation without adequate evidence or control for confounding factors. Assuming a direct cause-effect relationship without considering the complexity of economic systems could lead to biased interpretations.

Implementation: To gain insights into the factors that influence PCE:

(1) We sorted correlations based on their absolute values to identify indicators that exhibit the strongest linear relationships with PCE.

(2) The sorted correlations allowed us to view the magnitude and directionality of each relationship.

(3) We used a horizontal bar plot to depict the directional influence of each indicator on PCE.



Key Findings:

- Strong Positive Correlations: Service-providing industries, Total Nonfarm Employment, and Civilian Employment correlate highly with PCE, suggesting that employment levels 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, suggesting that

Team: Jan Nagtegaal | Aditya Sharma | Evan Trout

GitHub repository: https://github.com/janjannagtegaal/NowCasting-M1 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, indicating that these sectors contribute to consumer spending trends but to a lesser extent.

Implications:

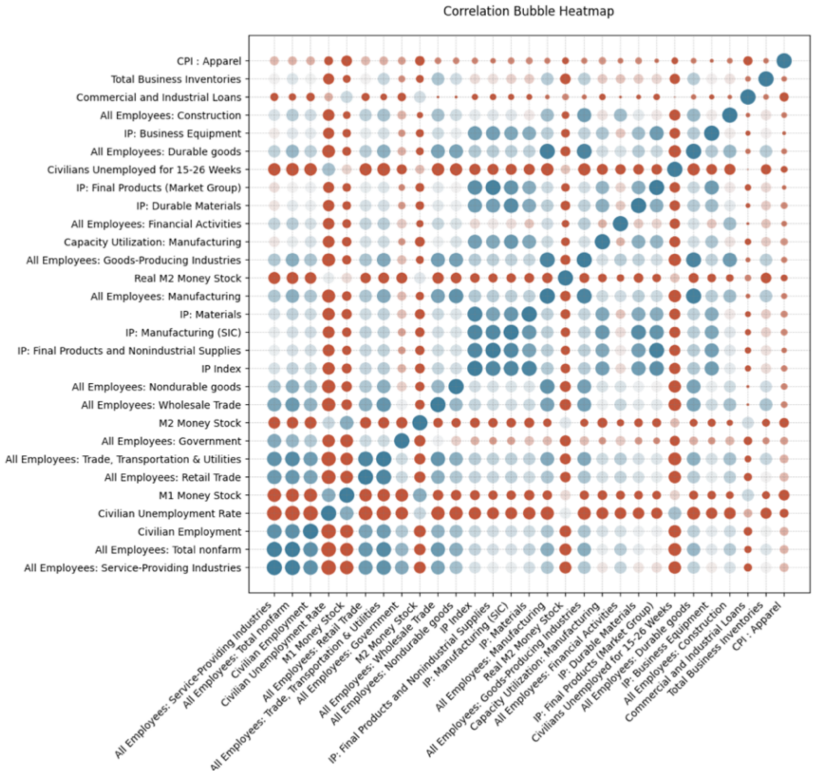
(1) The strong correlation between PCE and employment indicators highlights the importance of labour market health in driving consumer spending. This underscores the need for job creation and stability in service and non-farm sectors.
(2) The negative correlations of unemployment and money stock suggest that policies aimed at reducing unemployment and stabilising the money supply could positively influence PCE.
(3) While sectors like retail and government employment contribute to consumer spending, their impact may be influenced by broader economic conditions or specific sectoral trends, as indicated by moderate correlations.

Multicollinearity Analysis with Variance Inflation Factor (VIF)

Overview: Assessing the degree to which indicators are interrelated is essential to avoid distorting the actual relationship with PCE. High multicollinearity among variables can make it difficult to isolate the impact of individual indicators and lead to overfitting.

We use a two-pronged analytical approach to investigate multicollinearity among economic indicators: (1) Circular Correlation Heatmap visualisation and (2) Variance Inflation Factor (VIF) analysis. This helps us identify and address multicollinearity, refining our econometric modelling process.

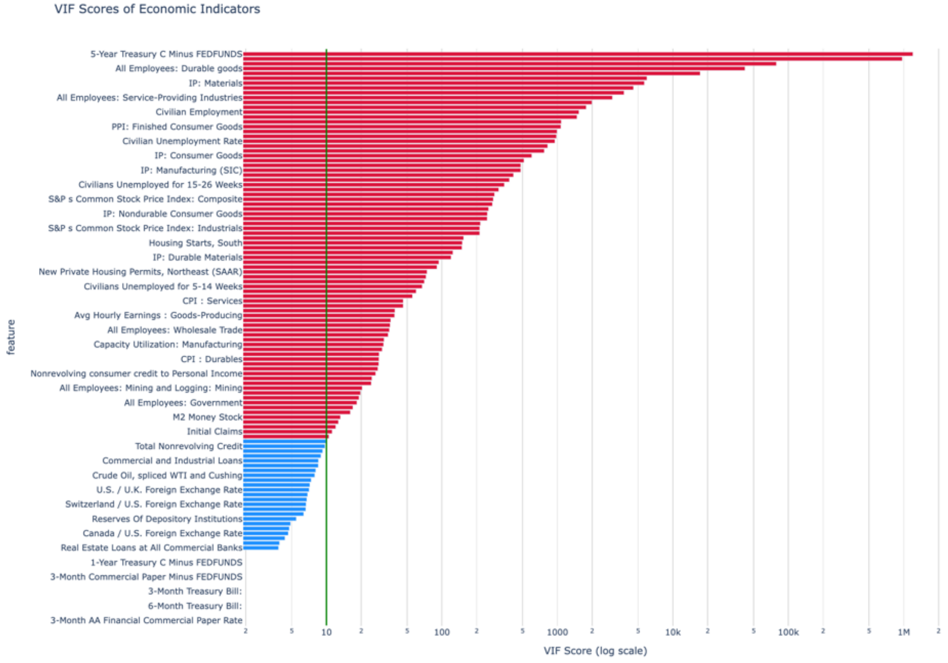
(1) Circular Correlation Heatmap Visualization: The Circular Correlation Heatmap provides an overarching view of indicator correlations, showcasing their interplay and the magnitude of their relationships. It adeptly highlights clusters of tightly correlated variables, allowing for the easy identification of potential multicollinearity among groups of indicators.



Through this visualisation, 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.

(2) VIF) analysis: This method quantifies the variance inflation of a regression coefficient due to predictor intercorrelations. A VIF above ten signals problematic multicollinearity and requires corrective measures in model specification.

Nowcasting Consumer Expenditure:



Observations: Several indicators of 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.

Next steps

(1) We will select proxies with strong correlations to PCE and evaluate their predictive strength through linear regression analysis, focusing on R^2 value.
(2) We will investigate seasonality and stationarity's impact on economic indicators.
(3) Select proxies based on their correlation and R^2 power over PCE
(4) Conduct further hypothesis testing on the linear relationship between indicators and PCE and the distribution of the residuals.
(5) Use Principal Component Analysis (PCA) for dimensionality reduction on our final proxies.

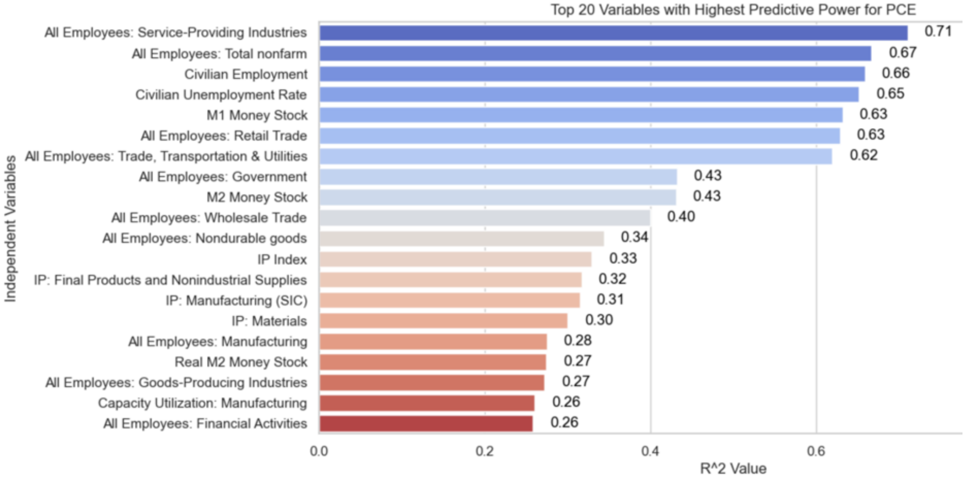
Linear Regression Analysis

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.

A higher R^2 value indicates a stronger linear relationship and more variance in PCE explained by the variable, earmarking it for further analysis as a significant driver of PCE.

Implementation

- Data Preparation: To ensure the integrity of our model, we exclude PCE from the independent variable pool and perform data cleansing by removing rows containing NaNs or infinite values.
- Fit a linear regression model to each selected variable and PCE to establish a baseline for analysis.
- Generate PCE predictions: using this model, for each independent variable, calculate the R^2 value to determine its explanatory power in PCE variations.



Team: Jan Nagtegaal | Aditya Sharma | Evan Trout

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

Nowcasting Consumer Expenditure:

Assessment of R2 Values:

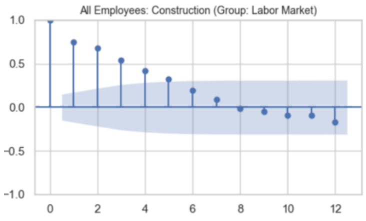
- Substantial Influence: Service-providing industries, Total Nonfarm Employees, and Civilian Employment have R^2 values > 0.65, indicating a considerable influence on PCE. Positive correlations affirm the pivotal role of employment sectors in driving consumer spending.
- Influence of Unemployment and Money Stock: Civilian Unemployment Rate and M1 Money Stock have negative correlations and moderate to high explanatory power, showing how changes in these areas can inversely impact PCE. These relationships underscore the sensitivity of consumer spending to broader economic conditions.
- Sector Diversity: PCE is correlated with multiple sectors, showcasing the complex interplay between economic activities and consumer spending.

Using indicators with higher R^2 values and correlation coefficients, we understand each variable's influence on PCE, identifying the most vital influencers and variables contributing unique insights into PCE dynamics.

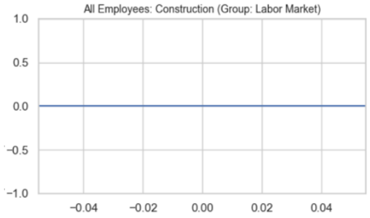
Seasonality Assessment:

The first critical step after selecting a subset of proxies involves assessing seasonality within our dataset. We identify indicators with notable seasonality by calculating Autocorrelation function (ACF) values for specified lags and identifying those exceeding a predetermined threshold. We adjust our series through seasonal decomposition by stripping away the seasonal component.

Before



After



Stationarity Assessment:

We ensure stationarity using the Augmented Dickey-Fuller (ADF) test to accurately model and forecast time series data. This test checks for a unit root in the series, indicating stationarity if the p-value is below 0.05. Economic datasets often diverge from a normal distribution, making this step crucial for further econometric modelling. Our selected indicators passed the ADF test, affirming their appropriateness for in-depth analysis.

Proxy Selection: After carefully considering the proxies, we have strategically and iteratively decided on a threshold to select our proxies for further analysis. The criteria:

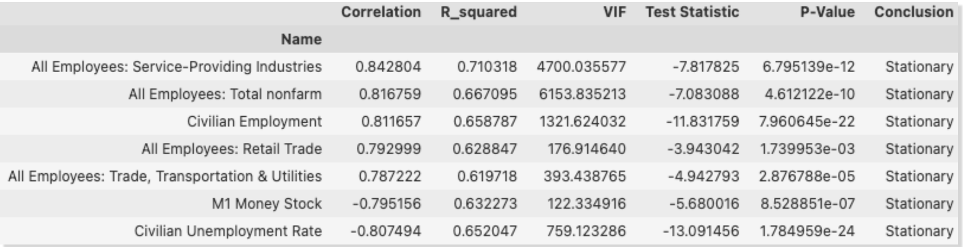
(1) have a correlation coefficient of 0.6

(2) R^2 score of 0.5

(3) be stationary

(4) not contain notable seasonality.

We have identified a refined list of 7 indicators that meet these criteria. 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).



	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

Nowcasting Consumer Expenditure:

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:

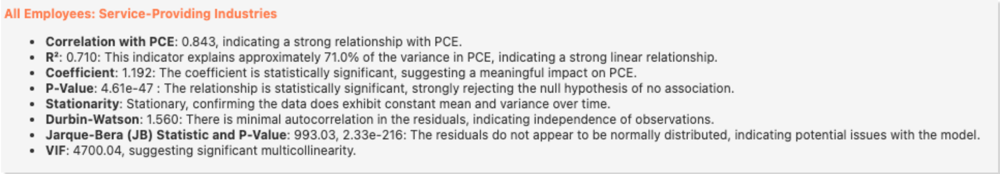
(1) Linear Relationship with PCE: The existence of a significant linear relationship with PCE:

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

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

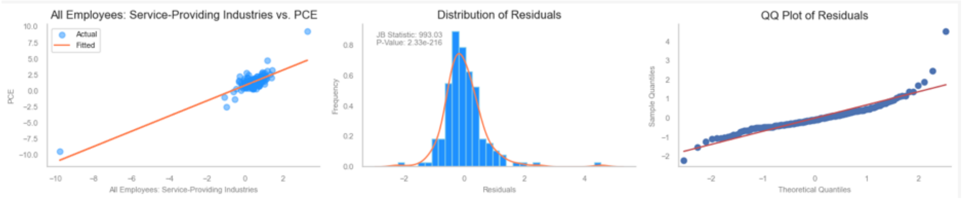
Null Hypothesis:	Alternative Hypothesis:
The residuals from the regression model are normally distributed, indicating the model's assumptions about the error term distribution are valid.	The residuals from the regression model are not normally distributed, suggesting potential issues with the model, 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.



Interpretation:

The chosen proxies show strong statistical significance and correlation with PCE, with:

1. A correlation coefficient above 0.78 and
2. R² values explaining 62.0% to 71.0% of the variance in PCE.
3. The statistical significance of these relationships is validated by p-values below 0.05, firmly rejecting the null hypothesis of no association.
4. However, there are issues with multicollinearity, as evidenced by high VIF scores and potential model assumption violations, highlighted by the Jarque-Bera test results.
5. Some instances of autocorrelation of residuals are also present, as indicated by the Durbin-Watson statistic. This poses challenges that require further model diagnostics and exploration of alternative modelling approaches or transformations to meet the assumptions of linear regression.

The relationship of these labour market indicators with PCE aligns with theoretical expectations. Employment levels and consumer spending are positively correlated, while the M1 Money Stock and Civilian Unemployment Rates negatively correlate with PCE. These correlations support economic theories that relate the money supply and unemployment rates inversely to consumer spending.

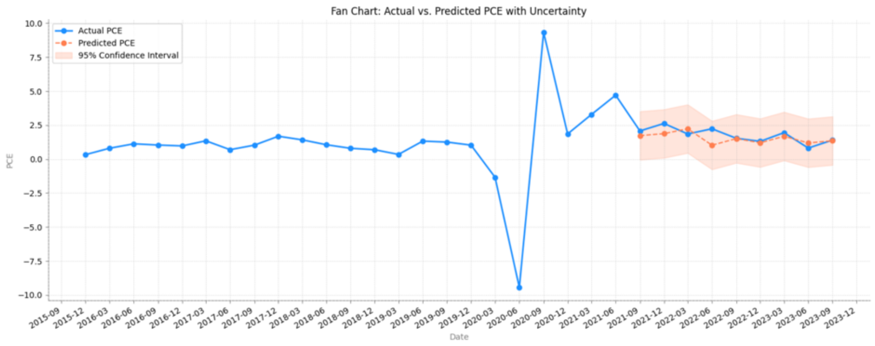
Principal Component Analysis (PCA) Analysis: Dimension Reduction

We implemented Principal Component Regression (PCR) analysis to refine our predictive models, combining Principal Component Analysis (PCA) with Linear Regression. This technique streamlines the dataset into principal components that can be used as new, uncorrelated predictors, potentially enhancing model

Nowcasting Consumer Expenditure:

performance and interpretability. We performed PCA, transforming the original variables into principal components used as new predictors. The PCR model was operationalised using a pipeline and running (1) StandardScaler for data normalisation, (2) PCA for dimensionality reduction, and (3) Linear Regression for prediction.

Regression Using Principal Components: We used two principal components as predictors in a Linear Regression model to forecast PCE. We combined actual PCE data with the model's predictions to contextualise our model's performance and associated uncertainty. This integration was visualised through a line chart, showing the model's forecasts alongside actual PCE values, with shading indicating the estimated prediction uncertainty.

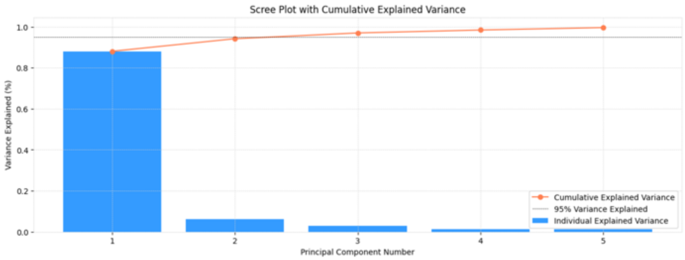


Accuracy Analysis: The model's predictive accuracy was measured using:

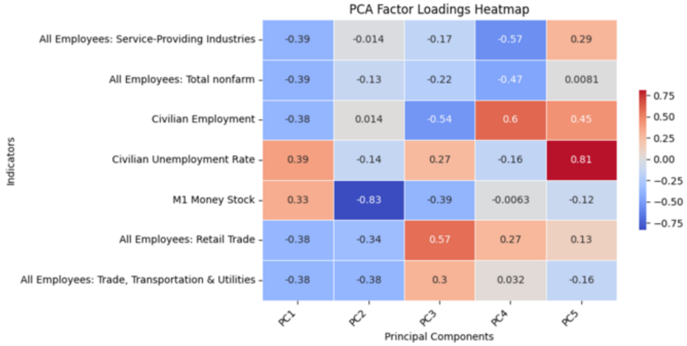
- Mean Absolute Error (0.3132) and
- Root Mean Squared Error (0.4069).
- Cross-validation performance was measured using the average MSE (0.4902).

The MAE and RMSE values indicate a relatively good accuracy in the model's predictions, while the low RMSE confirms its 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. Choosing the correct number of components is crucial to avoid overfitting. The Scree Plot showed a rapid decline in variance explained by successive principal components, indicating that the initial components captured the most significant portion of the variance.



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



Next Steps:

In the next project phase, we'll refine the model to improve accuracy and address statistical issues. We'll use advanced statistical tests and machine learning models like vector autoregression, dynamic factor models, Random Forest and Gradient Boosting. We'll create a real-time model with an API and dashboard for dynamic visualisation of predictions and actual economic indicators, reducing decision-making lag for policymakers and analysts.