**Real Time Inflation Index from dynamic data using web scarpping**

**Abstract**

The Consumer Price Index (CPI), which is a weighted average of prices for a range of products and services, shows a general increase in inflation. Products that are deemed characteristic of a particular consumption basket are taken into consideration when selecting the products that end up in the index. The index considers a number of variables, including the nation in question and the population's overall spending tendencies. The total CPI is determined by adding the weight of each individual item in the market basket.The utilization of prices obtained from online shops can serve as a valuable tool in the construction of high-frequency price indexes, thereby enhancing the existing official statistics. This study examines the capacity of India to align their inflation forecasts with government figures. The data were gathered from the primary online supermarkets. The aim of this study is to examine the different rates of inflation that different countries around the world experience and to publish our findings after the investigation is over. The principal objective is to construct an interactive dashboard that utilises a web-based user interface. The dashboard will be used to conduct research on various nations' rates of inflation.

The visualisation dashboard's user interface was designed to make it easy for users of different technical skill levels to understand the essential elements of the data that is being shown. It was designed to make it simple for users, regardless of technical expertise, to understand the data points that are deemed to be the most crucial. As a direct result of this endeavor, businesses will find it much easier to identify patterns and trends, which will improve their ability to make decisions.

**1. Introduction**

The accessibility of internet pricing data gives a distinctive prospect for the development of price indices and the assessment of inflation on a global scale. Special software has made it possible to acquire an exceptional quantity of micro-level price data by locating and consolidating comprehensive product information from various online sellers on the internet. The data gathering approach described can be conducted at far higher frequencies and at a significantly lower cost compared to existing methods. One of the primary applications of online price indexes is to acquire alternative inflation series in nations where official estimates have suffered a loss of trust, among its numerous possible uses.

The data collected for this study were obtained through daily online monitoring fetched from various online supermarkets like amazon, Bigbasket, etc. On a daily basis, an automated software systematically scanned the websites of prominent supermarkets, extracting detailed information about individual products and subsequently storing it in a centralised database. Over the course of time, a comprehensive panel database was created, containing intricate details pertaining to each product, such as prices, product IDs, and a categorical indicator. This study utilizes data from 3 supermarkets. On average, each retailer has distinct products.

In each country, an online price index is constructed by amalgamating online prices, employing conventional CPI methodology, and incorporating official category weights. Subsequently, every index is juxtaposed with a corresponding authoritative supermarket index, which is formulated exclusively based on the Consumer Price Index (CPI) components of Food, Beverages, and Household Products. These categories align with the ones accessible in the web data.

**2. Studies**

The scholarly literature pertaining to Consumer price Index (CPI) and real time inflation index in India is notably limited; yet, it has garnered the attention of researchers since the 1980s. Hence, we have done further studies to find out the scenario in India.

2.1 Online and Official Price Indexes: Measuring Argentina’s Inflation (May 3,2012) by Alberto Cavallo

The Author collected data from The Billion Prices Project (BPP) at MIT and used “web scraping” technique to collect prices for online commodities offered in major supermarkets in Argentina, Brazil, Chile, Colombia, and Venezuela from October 2007 to March 2011. He found that Matching is most effective at yearly frequencies and benefits from supermarkets with substantial market shares and representative cities in Brazil, Chile, Colombia, and Venezuela. Also, only one country, Argentina's internet inflation differs significantly from official estimates over time. The difference's magnitude and longevity are surprising. Internet inflation has outperformed official projections by 2–3 times for over 3.5 years. Internet index inflation averaged 20.14% per year, compared to 8.38% in official numbers. The difference was 65% in March 2011.

2.2 The use of online prices in the Norwegian Consumer Price Index (2015) by Ragnhild Nygaard

Ragnhild used web scraping technique from online sources such as data from e-commerce sites. This study has been done for Norway. Firstly, Data is collected from the online ecommerce sites for more than 100 of products. Then, Inflation Index is calculated for particular products and then compared with the offline product data CPI. He found that including price growth from various purchaser channels and correct depiction is crucial for maintaining relevance in price indices. Also, while our test estimates are confined to one consumer group, they suggest that internet costs may be highly erratic.

2.3 Consumer price index prediction using Long Short Term Memory (LSTM) based cloud computing (2019) by S Zahara, Sugiantoand M B Ilmiddaviq

The Authors had done the study for East Java province. Firstly, Data was collected from 34 basic food prices in 38 East Java cities. This data from 2014-2018 was gathered online from the Department of Industry and Commerce of East Java Province. And then Preprocessing of data is done and the model is being formed with LSTM technique and the algorithm is optimized. They found that most nations use CPI values to evaluate inflation as they accurately reflect economic conditions like food price fluctuations. Accurate predictions are necessary for handling unwanted situations and anticipating government actions. Nadam (Nesterov Adam), the current algorithm, performed best with an RMSE value of 4.088.

2.4 INFLATION HEDGING: A DYNAMIC APPROACH USING ONLINE PRICES (August 5, 2022) by Alberto Cavallo, Megan Czasonis, William Kinlaw, and David Turkington

The Authors had studied for the United States but this paper also analyzed data from 21 countries in total. They had collected data from online sources, specifically from the websites of multi-channel retailers in various countries. Web scraping technology has been used to gather price data from these online retailers. This data is then used to compute online inflation indices, which are used in the analysis to study the relationship between inflation and asset prices. They found that Online Inflation Predicts Official Inflation, Breakeven Inflation Predicts Future Inflation, Dynamic Inflation Hedging Strategy.

2.5 Real-Time Inflation Forecasting in a Changing World (Jan 28,2013) by Jan J. J. Groen,Richard Paap &Francesco Ravazzolo

This study has also been done for the United States. The data series used in the article are appropriately transformed to render them covariance stationary, and for those series that are prone to revisions, the original data vintages are provided. Bayesian Model Averaging (BMA) techniques has been used to forecast U.S. inflation using real-time data. BMA is a statistical method that considers multiple potential models or regressions and combines their predictions by assigning a probability distribution over models. This allows for model uncertainty to be taken into account in forecasting. They considered approximately 15 potential predictor variables along with lagged inflation rates to forecast quarterly inflation rates, specifically the Personal Consumption Expenditures (PCE) deflator and the Gross Domestic Product (GDP) deflator. The Authors found out Model Averaging and Structural Breaks, In-Sample properties like changes in inflation persistence and volatility over different time periods, Variable Selection like the models tend to select a smaller number of predictor variables, resulting in more parsimonious models, Density Forecasting by providing more precise forecasts for the tails of the inflation density distribution, particularly in the lower tails.

2.6 Annual forecasting of inflation rate in Iran: Autoregressive integrated moving average modeling approach (Jan 6,2021) by Samrad Jafarian-Namin, Seyyed Mohammad Taghi Fatemi Ghomi, Mohsen Shojaie, Saeed Shavvalpour

This study has been done for Iran. The data has been collected from the World bank. The Authors have used Box-Jenkins methodology and specifically focuses on using ARIMA (Autoregressive Integrated Moving Average) models to model and forecast the yearly inflation rate in Iran. They found out the significance of inflation forecasting in economic policy-making and decision-making processes. Accurate inflation forecasts can aid policymakers in assessing economic performance and implementing effective monetary policies.

2.7 Tracking and Modelling Prices Using Web-Scraped Price Microdata: Towards Automated Daily Consumer Price Index Forecasting (2018) by Ben Powell and Guy Nason

The main focus of the authors is to study the CPI in the United Kingdom. This study has 2 datasets : It includes daily web-scraped prices for 33 product categories from three prominent UK supermarkets' websites during 14 months.The second data set includes disaggregated CPI statistics for the same product categories and additional contributors, issued monthly by the ONS. The main methodology is to calculate empirical Bayes estimates for hyperparameters, the log-posterior density was optimized using the Nelder–Mead simplex approach. Then, they made Remarks on inferred consumer price index values followed by Remarks on inferred hyperparameters for product categories. They found that Estimates can identify goods that significantly impact aggregated CPI numbers and products with strong or weak correlations between web-scraped prices and survey prices. The CPI is a crucial indicator for economic and social users, as well as for official data like national accounts.

**3. Methodology**

The objective of this study is to analyze the Consumer Price Indices (CPI) for different categories, including food and beverages, clothing, housing, fuel and light, pan-tobacco-intoxicants, and miscellaneous items. The data has been collected using a combination of web scraping techniques and reliable sources available on the internet.

**I)Data Collection:**

A)Food and Beverages, Clothing:

* The process of gathering information on food and beverage, as well as clothing items, involved web scraping from popular e-commerce platforms such as Amazon and BigBasket. Web scraping is a technique that allows automated extraction of data from websites. In this context, it was utilized to collect details about various products, including their prices, descriptions, and customer reviews.
* Amazon and BigBasket Selection: These platforms were chosen due to their extensive product listings, diverse offerings, and popularity among consumers. Both platforms provide a wide range of products in the specified categories.
* Categories Defined: The focus was on two main categories: Food and Beverages, and Clothing. Within these categories, a variety of products were targeted to ensure a comprehensive dataset.
* Data Cleaning: Once the data was collected through web scraping, it underwent a thorough cleaning process to enhance its quality, reliability, and consistency. The key steps in data cleaning included:
* Handling Missing Values: Identify and address any instances where essential data points were not captured during scraping. This could involve imputing missing values or making decisions on how to handle the absence of certain information.
* Removing Duplicates: Eliminate redundant entries to ensure that each data point is unique. Duplicates might arise from repeated scraping or product variations.
* Standardizing Data Formats: Ensure consistency in the format of extracted data, including prices, descriptions, and reviews. This step helps in creating a standardized dataset that is easier to analyze.
* Dealing with Outliers: Identify and address any outliers or unusual data points that might skew the analysis. This step contributes to the overall accuracy of the dataset.
* Quality Assurance:Throughout the entire process, there was a focus on quality assurance to guarantee the reliability of the collected data. Rigorous checks were performed to validate the accuracy of information, and steps were taken to minimize potential biases or errors introduced during web scraping.
* Ethical Considerations:It's important to note that web scraping raises ethical considerations, and the process adhered to ethical guidelines. This includes respecting the terms of service of the targeted websites, ensuring user privacy, and avoiding any disruptive impact on the platforms being scraped.

B)Housing:

* Working Paper on Housing Price Indices: Data for housing was sourced from the working paper titled "477 Housing Price Indices in India."
* This document is a reliable source that provides comprehensive information on housing price indices.

C)Fuel and Light:

* Statista: Statista is recognized as a reliable statistical platform that aggregates data from diverse sources. Leveraging Statista adds a layer of credibility to the collected information, as it undergoes scrutiny for accuracy and reliability.
* Online Sources: In addition to Statista, other reputable online sources were consulted to ensure a well-rounded dataset. This approach enhances the diversity of data and minimizes dependency on a single platform, contributing to a more robust representation of fuel and light-related items.
* Components Included in Fuel and Light in CPI in India: Fuel ,Lighting ,Cooking Fuel ,Heating Fuels & Electricity

D)Pan, Tobacco, and Intoxicants:

* Statista and BigBasket: Information regarding pan, tobacco, and intoxicants was gathered from both Statista and BigBasket.
* Statista was used for historical and comparative data, while BigBasket provided real-time pricing information.

E)Miscellaneous:

* Online Reliable Sources: Data for miscellaneous items, including services like transport, education, and health, were collected from reliable online sources.
* These sources were carefully selected to ensure accuracy and completeness of the information.

**II)Data Comparison and Analysis:**

A) CPI Calculation:

* The collected data underwent a meticulous process to compute the Consumer Price Index (CPI) for each category.
* The CPI serves as a vital metric, illustrating the average fluctuation in prices over time for a predetermined basket of goods and services.
* By leveraging the gathered data, the CPI calculation provides a quantitative measure of the inflationary or deflationary trends within specific categories, offering valuable insights into economic dynamics.

B) Offline CPI Value:

* A comparative analysis was conducted by juxtaposing the computed CPI values with the offline CPI data available in our repository.
* This offline CPI dataset served as a pivotal benchmark, allowing for a rigorous validation of the accuracy and reliability of the web-scraped and sourced data.
* By ensuring alignment with established offline CPI values, the credibility and fidelity of the collected information were substantiated, fortifying the overall robustness of the CPI calculations.

C) Statistical Analysis:

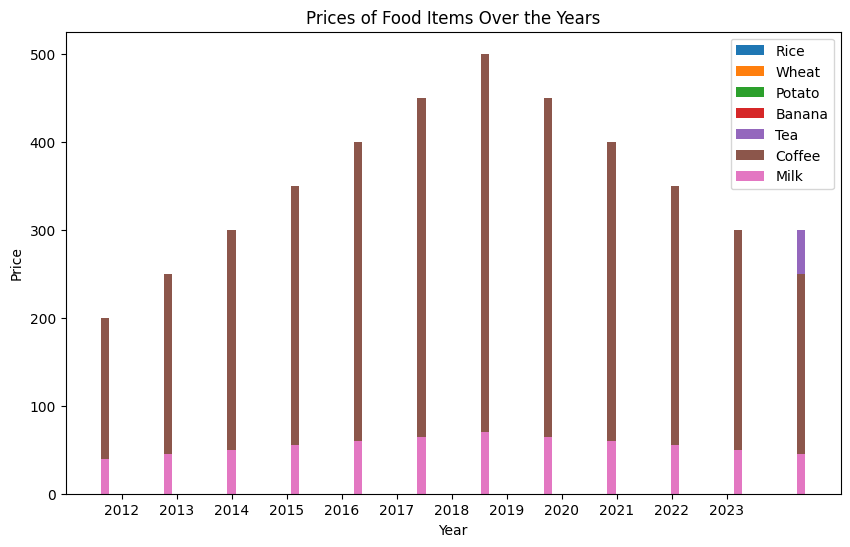
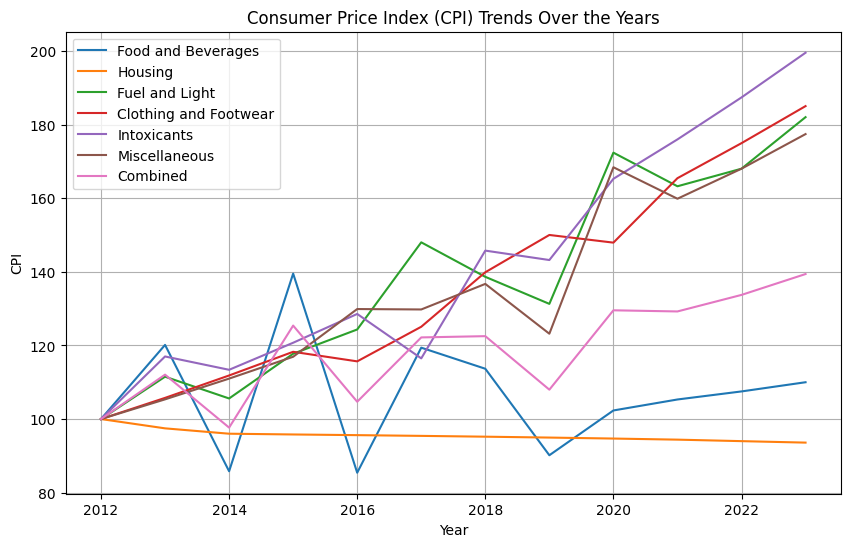
* To unveil deeper insights into the distribution and characteristics of the CPI values for each category, a comprehensive statistical analysis was performed.
* This involved the computation of key descriptive statistics, including mean, median, and standard deviation.
* These statistical measures provided a nuanced understanding of the central tendency and variability within the CPI datasets.
* Such analysis goes beyond mere numerical values, offering a richer comprehension of the distributional characteristics inherent in the price movements of the selected goods and services.

D) Graphical Representation:

* The findings were vividly communicated through graphical representation, employing charts and graphs to convey a visual narrative of the trends and variations present in the CPI values across diverse categories.
* This graphical approach enhances interpretability, enabling stakeholders to discern patterns, anomalies, and trends at a glance.
* Whether through line charts illustrating temporal trends or bar graphs highlighting comparative variations, the visual representation fosters a clearer comprehension of the intricate dynamics encapsulated within the CPI data, thereby facilitating more informed decision-making.

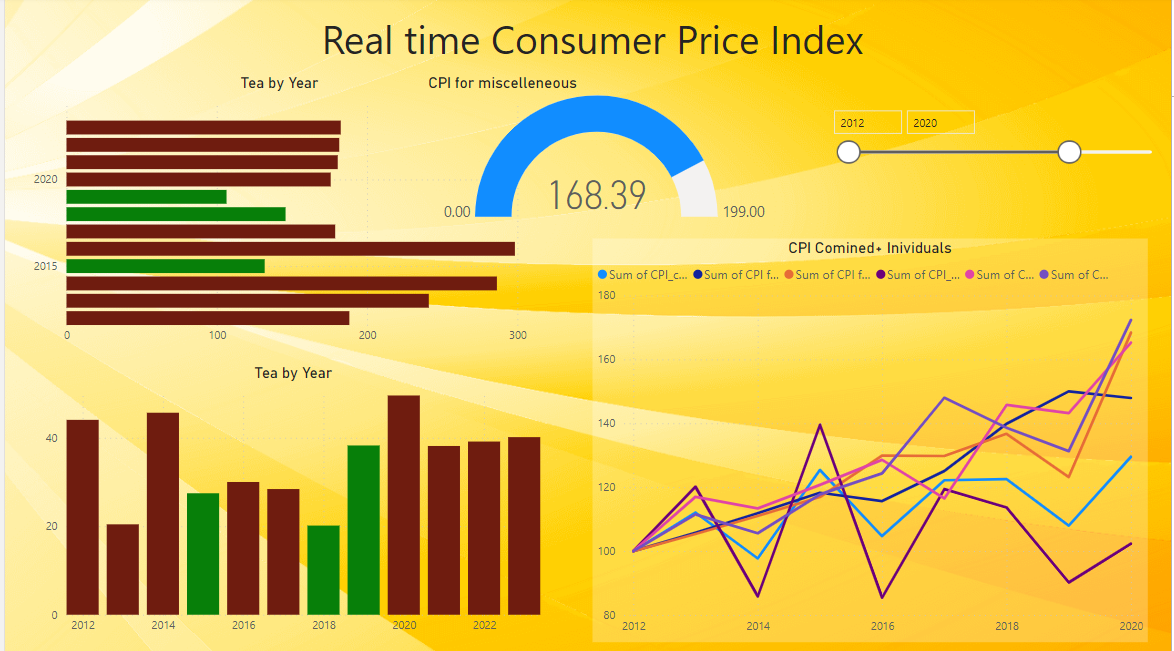
**4. Results**

**CPI trends over years**

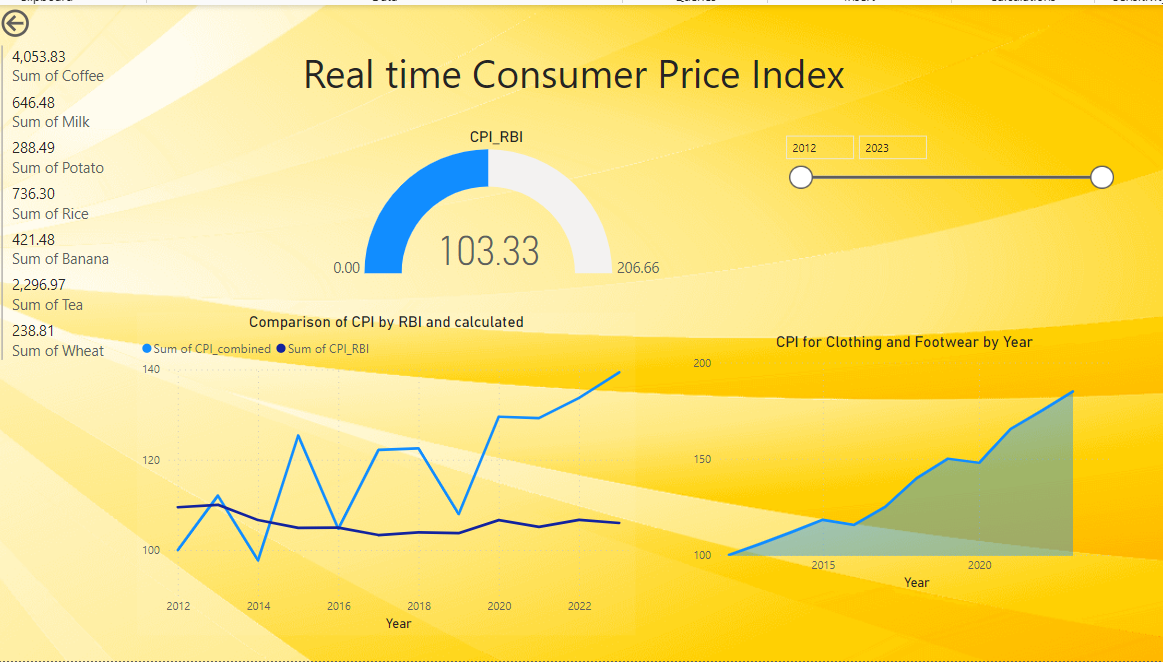


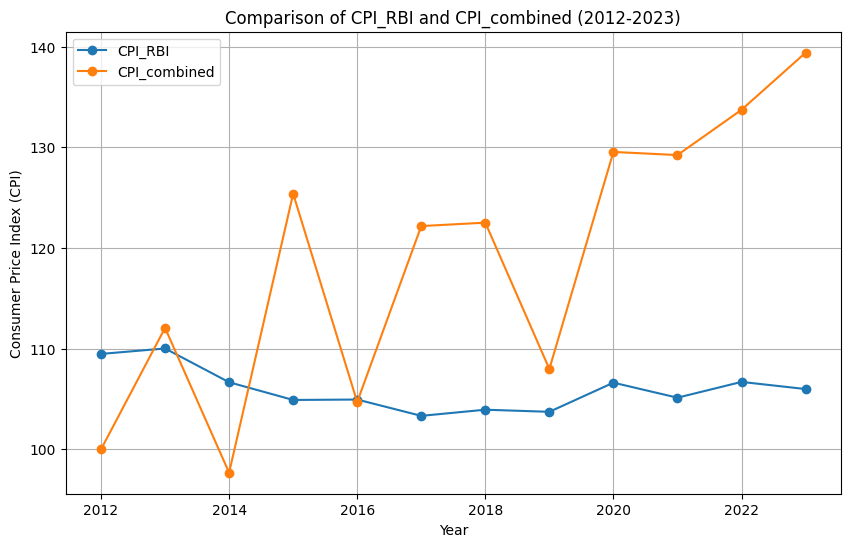
**Dashboard image**

**Page 1**

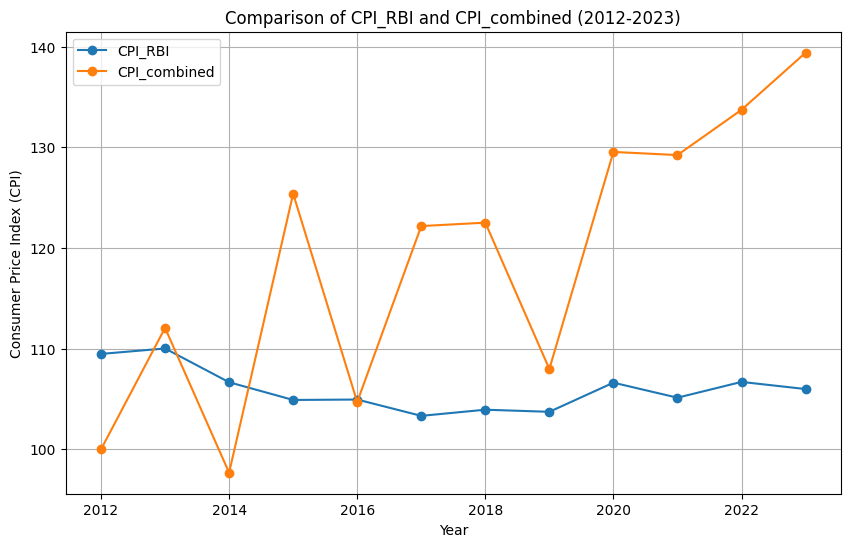
****

**Page 2**

****



| Year | CPI\_RBI | CPI\_combined |
| --- | --- | --- |
| 2012 | 109.4789969 | 100 |
| 2013 | 110.0178785 | 112.06316 |
| 2014 | 106.6656567 | 97.68098049 |
| 2015 | 104.9069734 | 125.3888287 |
| 2016 | 104.9482163 | 104.7035686 |
| 2017 | 103.3281734 | 122.1741892 |
| 2018 | 103.9388265 | 122.5148174 |
| 2019 | 103.7295057 | 107.9819616 |
| 2020 | 106.6234368 | 129.5361903 |
| 2021 | 105.1314075 | 129.2258255 |
| 2022 | 106.6990341 | 133.7183256 |
| 2023 | 105.99 | 139.3886 |



**5. Conclusion**

In conclusion, our study underscores the significance of dynamic data sources in obtaining more comprehensive Consumer Price Index (CPI) values compared to the traditional indices released by the Reserve Bank of India (RBI). The limitations of relying solely on RBI's CPI calculation methodology are evident, as it often excludes numerous products widely used in our daily lives. By leveraging the vast and continuously evolving landscape of the internet, we accessed a broader spectrum of products, capturing the dynamic nature of consumer behavior more accurately. The inclusion of these additional data sources contributes to a more nuanced and representative understanding of inflation trends, enabling a more informed economic analysis. As we navigate an increasingly interconnected world, harnessing diverse data streams becomes imperative for refining the accuracy and relevance of economic indicators like CPI.

**6. References**

1. Jaworski, K. (2021). Measuring food inflation during the COVID-19 pandemic in real time using online data: a case study of Poland. *British Food Journal*, *123*(13), 260-280.
2. Aparicio, D., & Bertolotto, M. I. (2020). Forecasting inflation with online prices. *International Journal of Forecasting*, *36*(2), 232-247.
3. Budiastuti, I. A., Nugroho, S. M. S., & Hariadi, M. (2017, July). Predicting daily consumer price index using support vector regression method. In *2017 15th international conference on quality in research (QiR): international symposium on electrical and computer engineering* (pp. 23-28). IEEE.
4. Powell, B., Nason, G., Elliott, D., Mayhew, M., Davies, J., & Winton, J. (2018). Tracking and modelling prices using web-scraped price microdata: towards automated daily consumer price index forecasting. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *181*(3), 737-756.
5. Zahara, S., & Ilmiddaviq, M. B. (2020). Consumer price index prediction using Long Short Term Memory (LSTM) based cloud computing. In *Journal of Physics: Conference Series* (Vol. 1456, No. 1, p. 012022). IOP Publishing.
6. Nygaard, R. (2015). The use of online prices in the Norwegian Consumer Price Index. *Statistics Norway*, 1-16.
7. Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. *Journal of Monetary Economics*, *60*(2), 152-165.
8. Cavallo, A., Czasonis, M., Kinlaw, W. B., & Turkington, D. (2022). Inflation Hedging: A Dynamic Approach Using Online Prices. In *Inflation Hedging: A Dynamic Approach Using Online Prices: Cavallo, Alberto| uCzasonis, Megan| uKinlaw, William B.| uTurkington, David*. [Sl]: SSRN.
9. Groen, J. J., Paap, R., & Ravazzolo, F. (2013). Real-time inflation forecasting in a changing world. *Journal of Business & Economic Statistics*, *31*(1), 29-44.
10. Jafarian‐Namin, S., Fatemi Ghomi, S. M. T., Shojaie, M., & Shavvalpour, S. (2021). Annual forecasting of inflation rate in Iran: Autoregressive integrated moving average modeling approach. *Engineering Reports*, *3*(4), e12344.
11. Garcia, M. G., Medeiros, M. C., & Vasconcelos, G. F. (2017). Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting*, *33*(3), 679-693.
12. Macias, P., Stelmasiak, D., & Szafranek, K. (2023). Nowcasting food inflation with a massive amount of online prices. *International Journal of Forecasting*, *39*(2), 809-826.
13. Adubisi, O. D., David, I. J., James, F. E., Awa, U. E., & Terna, A. J. (2018). A predictive Autoregressive Integrated Moving Average (ARIMA) Model for forecasting inflation rates. *Res. J. Bus. Econ. Manag*, *1*(1), 1-8.
14. Horrell, S. (2023). Household consumption patterns and the consumer price index, England, 1260–1869. *The Economic History Review*.
15. Ejukwa, J. O., Essi, I. D., & Dum, D. Z. Autoregressive Integrated Moving Average Modeling of Nigeria Inflation Rate Before and within the Early Part of Covid-19 Era.
16. Uriarte, J. I., Ramírez Muñoz de Toro, G. R., & Larrosa, J. (2019). Web scraping based online consumer price index: The “IPC Online” case. *Journal of Economic and Social Measurement*, *44*(2-3), 141-159.
17. Ericsson, N. R., Irons, J. S., & Tryon, R. W. (2001). Output and inflation in the long run. *Journal of Applied Econometrics*, *16*(3), 241-253.
18. Dunn, A., Grosse, S. D., & Zuvekas, S. H. (2018). Adjusting health expenditures for inflation: a review of measures for health services research in the United States. *Health services research*, *53*(1), 175-196.