

**INTRODUCTION TO DATA MANAGEMENT PROJECT REPORT**

(Project Semester August-December 2018)

***STATE LEVEL CONSUMER PRICE INDEX (RURAL/URBAN)***

***(USAGE OF EXCEL)***

Submitted by

Bharath Simha Reddy Maram

Registration No: 11602281

Programme: Integrated B. Tech – MTech

Section: KEM42

Course Code: INT217

Under the Guidance of

**SAMEER FAROOQ**

**Discipline of CSE/IT**

**Lovely School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Bharath Simha Reddy Maram bearing Registration no. 11602281 has completed INT 217 project titled, **“State Level Consumer Price Index(Rural/Urban) (Usage of EXCEL)”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, Mangena Bharath Simha Reddy Maram, student of Integrated B. Tech – M.Tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature:

Registration No. 11602281 Name of the student: Bharath Simha Reddy M

Acknowledgement

*I am grateful to my instructor Mr. Sameer Farooq. He has truly been a great source of inspiration, constructive criticism, insight and input. The knowledge he has imparted and the patience they have displayed was vital in completing this study. I am thankful for the immense encouragement and morale support that he graciously provided during this study.*

*I owe much of my academic and personal success to my parents, who, by example, provided me with the motivation and courage to pursue this course in the field of my interest. Special thanks to all my friends, near and far, for their love and support that made me finish my Project and its report successfully.*

BHARATH SIMHA REDDY MARAM

**Table of Contents**

**Introduction . . . . . . . . . . 1**

*Significance of CPI . . . . . . . . 1*

*How CPI calculated in India? . . . . . . 2*

*CPI calculation process . . . . . . . 2*

**Scope of Analysis . . . . . . . . . 3**

**Source of Dataset . . . . . . . . . 3**

**ETL Process . . . . . . . . . 4**

*What is ETL? . . . . . . . . . 4*

*Why do we need ETL? . . . . . . . . 4*

*ETL process in Data Warehouse. . . . . . . 5*

*Modern ETL Process . . . . . . . . 8*

*ETL Vs ELT . . . . . . . . . 9*

**Analysis on Dataset . . . . . . . . . 10**

Description . . . . . . . . . 10

Concepts of Excel used in Analysing Data . . . . . 10

Consumer Price Index calculated on basis of Sectors . . . 11

Rural Sector . . . . . . . . 11

Urban Sector . . . . . . . . 13

Rural Urban Sector . . . . . . . 15

Consumer Price Index calculated on basis of State-wise . . . 16

**Bibliography . . . . . . . . . . 18**

**Introduction**

A Consumer Price Index (CPI) measures changes in the price level of market basket of consumer goods and services purchased by households. The CPI is a statistical estimate constructed using the prices of a sample of representative items whose prices are collected periodically.

It is one of several price indices calculated by most national statistical agencies. The annual percentage change in a CPI is used as a measure of inflation. A CPI can be used to index (i.e. adjust for the effect of inflation) the real value of wages, salaries, pensions, for regulating prices and for deflating monetary magnitudes to show changes in real values. In most countries, the CPI, along with the population census, is one of the most closely watched national economic statistics.

***The Significance of CPI***

The index is usually computed monthly, or quarterly in some countries, as a weighted average of sub-indices for different components of consumer expenditure, such as food, housing, shoes, clothing, each of which is in turn a weighted average of sub-sub-indices. These indices compare prices each month with prices in the price-reference month. The weights used to combine them into the higher-level aggregates, and then into the overall index, relate to the estimated expenditures during a preceding whole year of the consumers covered by the index on the products within its scope in the area covered.

The coverage of the index may be limited. Consumers' expenditure abroad is usually excluded; visitors' expenditure within the country may be excluded in principle if not in practice; the rural population may or may not be included; certain groups such as the very rich or the very poor may be excluded. Saving and investment are always excluded, though the prices paid for financial services provided by financial intermediaries may be included along with insurance.

***How CPI is calculated in India?***

CPI in India is calculated by taking a basket of 299 commodities as compared to 676 commodities in Wholesale Price Index (WPI). Basically, CPI is calculated by considering the retail price change of goods and services and by taking the average weighted value of each item in the basket.

***The CPI Calculation Process***

Calculating Consumer Price Index (and the inflation rate) follows a four-step process: 1) Fixing the market basket

2) calculating the basket’s cost

3) computing the index

4) computing the inflation rate.

The CPI market basket represents all goods and services that are purchased for consumption by a specific reference population (e.g. the urban population within the US). It includes items from more than 200 categories and eight major groups:

1.Food and Beverages

2.Housing

3.Apparel

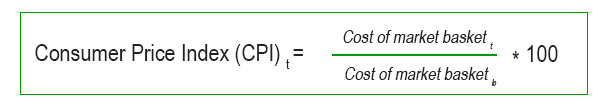
4.Transportation

5.Medical Care

6.Recreation

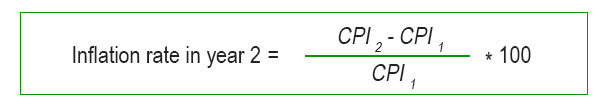
7.Education

8.Communication

9.Other Goods and Services

Cost of market basket t t🡪 In the given year

Cost of market basket b b🡪 In the base year



***Scope of Analysis:***

The scope of the Consumer Price Index (CPI) is defined to indicate what the CPI is intended to measure. Since there are many uses of the CPI, its scope has been defined to suit as many purposes as possible. The coverage of the index may be limited. Consumers expenditure abroad is usually excluded; visitors' expenditure within the country may be excluded in principle if not in practice; the rural population may or may not be included; certain groups such as the very rich or the very poor may be excluded. Saving and investment are always excluded, though the prices paid for financial services provided by financial intermediaries may be included along with insurance.

**Source of Dataset**

Open Government Data (OGD) Platform India or <data.gov.in> is a platform for supporting Open Data initiative of Government of India. This portal is single point access to datasets, documents, services, tools and applications published by ministries, departments, and organisations of the Government of India. It combines and expands the best features of India government’s India.gov.in and the US government’s <data.gov> project. So, the data obtained from <data.gov.in> is most reliable data and I preferred this site to get my data.

Dataset Reference: <https://data.gov.in/resources/state-level-consumer-price-index-ruralurban-december-2017-base-2012100>

**ETL Process**

***What is ETL?***

ETL is an abbreviation of Extract, Transform, and Load. In this process, an ETL tool extracts the data from different RDBMS source systems then transforms the data like applying calculations, concatenations, etc. and then load the data into the Data Warehouse system.

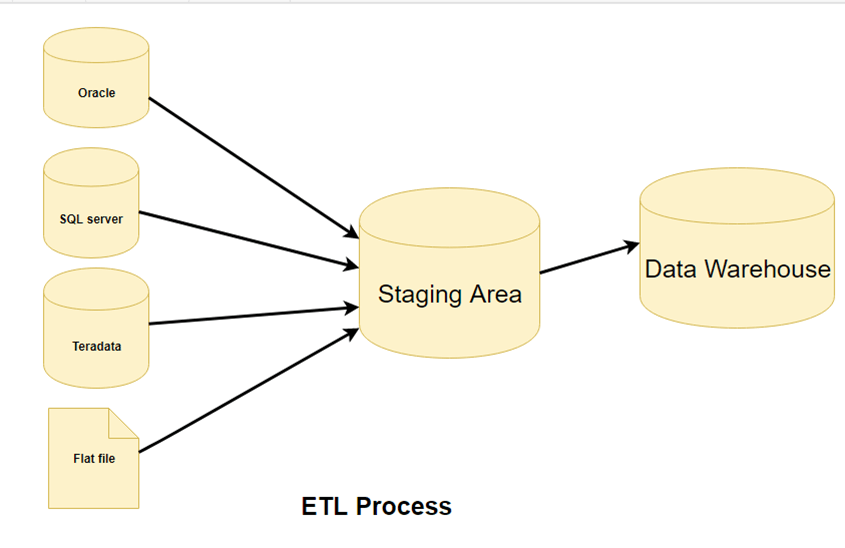
It’s tempting to think a creating a Data Warehouse is simply extracting data from multiple sources and loading into database of a Data Warehouse. This is far from the truth and requires a complex ETL process. The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives and is technically challenging.

In order to maintain its value as a tool for decision makers, Data Warehouse system needs to change with business changes. ETL is a recurring activity (daily, weekly, monthly) of a Data Warehouse system and needs to be agile, automated, and well documented.

***Why do you need ETL?***

There are many reasons for adopting ETL in the organization:

* It helps companies to analyse their business data for taking critical business decisions.
* Transactional databases cannot answer complex business questions that can be answered by ETL.
* A Data warehouse provides a common data repository.
* ETL provides a method of moving the data from various sources into a data warehouse.
* As data sources change, the Data Warehouse will automatically update.
* Well-designed and documented ETL system is almost essential to the success of a Data Warehouse project.
* Allow verification of data transformation, aggregation and calculations rules.
* ETL process allows sample data comparison between the source and the target system.
* ETL process can perform complex transformations and requires the extra area to store the data.
* ETL helps to Migrate data into a Data Warehouse. Convert to the various formats and types to adhere to one consistent system.
* ETL is a predefined process for accessing and manipulating source data into the target database.
* ETL offers deep historical context for the business.
* It helps to improve productivity because it codifies and reuses without a need for technical skills.



***ETL Process in Data Warehouses***

ETL is a 3-step process:

Step – 1: Extraction

Step – 2: Transformation

Step – 3: Loading

***Step – 1: Extraction***

In this step, data is extracted from the source system into the staging area. Transformations if any are done in staging area so that performance of source system in not degraded. Also, if corrupted data is copied directly from the source into Data Warehouse database, rollback will be a challenge. Staging area gives an opportunity to validate extracted data before it moves into the Data Warehouse.

Data Warehouse needs to integrate systems that have different DBMS, Hardware, Operating Systems and Communication Protocols. Sources could include legacy applications like Mainframes, customized applications, point of contact devices like ATM, Call switches, text files, spreadsheets, ERP, data from vendors, partners amongst others

Hence one needs a logical data map before data is extracted and loaded physically. This data map describes the relationship between sources and target data.

***Three Data Extraction methods***

1. Full Extraction
2. Partial Extraction – without update notification.
3. Partial Extraction – with update notification

Irrespective of the method used, extraction should not affect performance and response

time of the source systems. These source systems are live production databases. Any slow down or locking could affect company’s bottom line.

***Some validations are done during Extraction***

* Reconcile records with the source data
* Make sure that no spam/unwanted data loaded
* Data type check
* Remove all types of duplicate/fragmented data
* Check whether all the keys are in place or not

***Step – 2: Transformation***

Data extracted from source server is raw and not usable in its original form. Therefore, it needs to be cleansed, mapped and transformed. In fact, this is the key step where ETL process adds value and changes data such that insightful BI reports can be generated.

In this step, you apply a set of functions on extracted data. Data that does not require any transformation is called as **Direct move** or **Pass through data.**

In transformation step, you can perform customized operations on data. For instance, if the user wants sum-of-sales revenue which is not in the database. Or if the first name and the last name in a table is in different columns. It is possible to concatenate them before loading.

***Validations are done during this stage***

* Filtering – Select only certain columns to load
* Using rules and lookup tables for Data standardization
* Character set conversion and encoding handling
* Conversion of units of measurements like Date Time Conversion, Currency conversions, Numerical Conversions, etc.
* Data threshold validation check. For example, age cannot be more than two digits.
* Data flow validation from the staging area to the intermediate tables.
* Required fields should not be left blank.
* Cleaning (for example, mapping NULL to 0 or Gender Male to “M” and Female to “F” etc.)
* Split a column into multiples and merging multiple columns into a single column.
* Transposing rows and columns
* Use lookups to merge data
* Using any complex data validation (e.g., if the first two columns in a row are empty then it automatically rejects the row from processing)

***Step – 3: Loading***

Loading data into the target data warehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period. Hence, load process should be optimized for performance.

In case of load failure, recover mechanisms should be configured to restart from the point of failure without data integrity loss. Data Warehouse admins need to monitor, resume, cancel loads as per prevailing server performance.

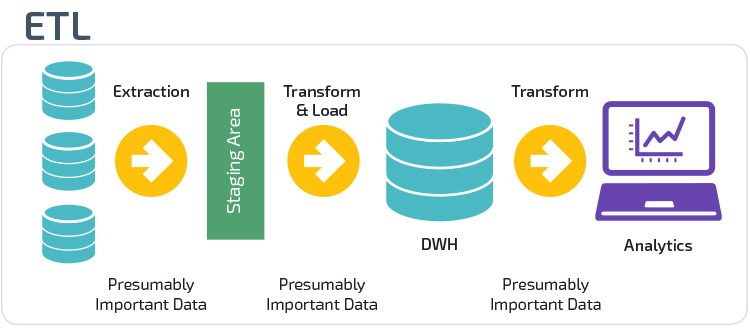
***Types of Loading***

* **Initial Load** – Populating all the Data Warehouse tables
* **Incremental Load** – Applying ongoing changes as when needed periodically.
* **Full Refresh** – Erasing the contents of one or more tables and reloading with fresh data.

***Load Verification***

### Ensure that the key field data is neither missing nor null.

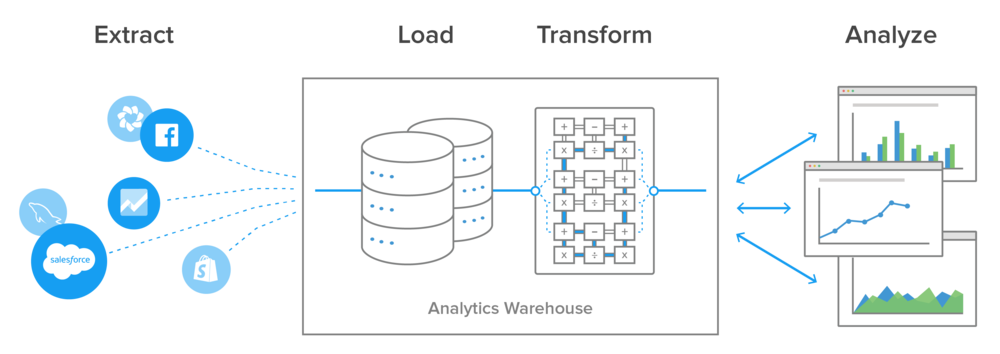
* Test modelling views based on the target tables.
* Check that combined values and calculated measures.
* Data checks in dimension table as well as history table.
* Check the BI reports on the loaded fact and dimension table.



***Modern ETL Process***

Modern technology has changed most organizations approach to ETL, for several reasons. The biggest is the advent of powerful analytics warehouse like Amazon Redshift and Google BigQuery. These newer cloud-based analytics databases have the horsepower to perform transformations in place rather than requiring a special staging area.

Also, data today is frequently analysed in raw form rather than from preloaded OLAP summaries. This has led to the development of lightweight, flexible, and transparent ETL systems with processes that look something like this:



The biggest advantage to this setup is the transformations and data modelling happen in the analytics database. This gives the BI team, data scientist, and analysts greater control over how they work with it, in a common language they all understand.

***ETL Vs ELT***

|  |  |  |
| --- | --- | --- |
| **Difference between ETL Vs ELT** | | |
| **Parameters** | **ETL** | **ELT** |
| **Process** | Data is transformed at staging server and then transferred to Data Warehouse DB | Data remains in the DB of the Data Warehouse |
| **Code Usage** | Used for Compute Intensive Transformation Small amount of data | Used for High amounts of data |
| **Transformation** | Transformations are done in ETL server/staging area. | Transformations are performed in the target system |
| **Time-Load** | Data first loaded into staging and later loaded into target system. Time intensive. | Data loaded into target system only once. Faster. |
| **Time-Transformation** | ETL process needs to wait for transformation to complete. As data size grows, transformation time increases. | In ELT process, speed is never dependant on the size of the data. |
| **Time- Maintenance** | It needs highs maintenance as you need to select data to load and transform. | Low maintenance as data is always available. |
| **Implementation Complexity** | At an early stage, easier to implement. | To implement ELT process organization should have deep knowledge of tools and expert skills. |
| **Support for Data warehouse** | ETL model used for on-premises, relational and structured data. | Used in scalable cloud infrastructure which supports structured, unstructured data sources. |
| **Complexity** | The ETL process loads only the important data, as identified at design time. | This process involves development from the output-backward and loading only relevant data. |
| **Cost** | High costs for small and medium businesses. | Low entry costs using online Software as a Service Platforms. |
| **Lookups** | In the ETL process, both facts and dimensions need to be available in staging area. | All data will be available because Extract and load occur in one single action. |
| **Calculations** | Overwrites existing column or Need to append the dataset and push to the target platform. | Easily add the calculated column to the existing table. |
| **Hardware** | Most tools have unique hardware requirements that are expensive. | Being Saas hardware cost is not an issue. |
| **Support for Unstructured Data** | Mostly supports relational data | Support for unstructured data readily available. |

***ANALYSIS ON DATASET***

***Description:***

Dataset can be analysed into two major divisions, mentioned as:

1. Consumer Price Index calculated on basis of Sectors
2. CPI of States based on Rural sector
3. CPI of States based on Urban sector
4. CPI of States based on Rural Urban sector
5. Consumer Price Index calculated on basis of State-wise from 2011-2017

In the first case, we divide the given data into 3 sub-data parts as Rural, Urban, Rural Urban, which shows the index values of different states in different sectors.

In the second case, we analyse the data based on states, which shows in particular state we can compare the index values of different sectors in each state.

***Concepts of Excel used in Analysing the Data***

* Pivot Tables
* Pivot Charts
* Statistical Functions
  + Maximum
  + Minimum
  + Mean
* Linking of different sheets in Excel
* Dashboard

***1.*** ***Consumer Price Index calculated on basis of Sectors***

CPI has two components, one a representative of the entire urban population, viz. CPI (Urban), and another for the entire rural population, viz. CPI (Rural) These indices reflect the changes in the price levels of various goods and services consumed by the urban and rural population respectively. The indices are compiled at State/UT and all-India levels and are based on 2010 as base year. CPI (urban) covers 310 towns while the span of CPI(rural) is 1181 villages. Index Numbers for both rural and urban areas and combined have been started from January 2011 index onwards. Provisional indices based on the data available are first released with the time lag of 30 days. Revised and final numbers with complete data for all India and for all the States/UTs will be released with a time lag of two months.

From the given dataset, we divide the Data into three sectors:

1. Rural Sector
2. Urban Sector
3. Rural Urban Sector

***Rural Sector:***

Rural areas are also known as the 'countryside' or a 'village' in India. It has a very low population density. In rural areas, agriculture is the chief source of livelihood along with fishing, cottage industries, pottery etc.

According to the Planning Commission, a town with a maximum population of 15,000 is considered rural in nature. In these areas the panchayat makes all the decisions.

Dataset can be divided into Rural Sector-Dataset in which Charts are formed for all the States and Index of each states for specified year as follows:

From this Graph, we know from the Rural data that on X-axis: Year-wise and on Y-axis: Index values. Different colours show the average index of different states and union territories in each year.

In Rural Sector, the highest Index record is Meghalaya-161.67

In Rural Sector, the lowest Index record is Telangana-0

In Telangana state, there is no data / elevation in graph, as state not formed in the years 2011,2012,2013. From 2014-2017 we can see elevation in graph.

As we can’t conclude the index values properly, this graph again simplified for single state as follows:

In this pattern, further states can also be drawn.

***Urban Sector:***

An Urban area is a human settlement with high population densityand infrastructure of built environment. Urban areas are created through urbanization and are categorized by urban morphology as cities, towns, suburbs.

In 2014 there were 7.2 billion people living on the planet, of which the global urban population comprised 3.9 billion.

Dataset can be divided into Urban Sector-Dataset in which Charts are formed for all the States and Index of each states for specified year as follows:

From this Graph, we know from the Urban data that on X-axis: Year-wise and on Y-axis: Index values. Different colours show the average index of different states and union territories in each year.

In Urban Sector, the highest Index record is Tripura-145.58

In Urban Sector, the lowest Index record is Arunachal Pradesh-0

In Arunachal Pradesh state, there is no Urban areas. So, no elevation in graph.

As we can’t conclude the index values properly, this graph again simplified for single state as follows:

In this pattern, further states can also be drawn.

***Rural Urban Sector:***

Rural Urban Sector is the average Index of Rural and Urban sectors which data is provided in the Dataset.

Dataset can be divided into Rural Urban Sector-Dataset in which Charts are formed for all the States and Index of each states for specified year as follows:

From this Graph, we know from the Rural Urban data that on X-axis: Year-wise and on Y-axis: Index values. Different colours show the average index of different states and union territories in each year.

In Rural Urban Sector, the highest Index record is Meghalaya-154.19

In Rural Urban Sector, the lowest Index record is Telangana and Arunachal Pradesh-0

In Telangana state, there is no data / elevation in graph, as state not formed in the years 2011,2012,2013. From 2014-2017 we can see elevation in graph.

In Arunachal Pradesh state, there is no Urban areas. So, no elevation in graph.

As we can’t conclude the index values properly, this graph again simplified for single state as follows:

In the above pattern, further states can also be drawn.

***2. Consumer Price Index calculated on basis of State-wise from 2011-2017***

From the dataset, we can analyse the data based on state. We can compare the two or more states at a time and analyse the index difference between different sectors in them. So, we can easily identify the highest index and lowest index of each state in each sector.

From this graph, we can conclude the index values of different sectors in different states.

On X-axis: States and on Y-axis: Index values.

Highest Index record is Meghalaya-Rural

Lowest Index recorded is Arunachal Pradesh-Urban, Telangana

In Telangana state, there is no data / elevation in graph, as state not formed in the years 2011,2012,2013. From 2014-2017 we can see elevation in graph.

In Arunachal Pradesh state, there is no Urban areas. So, no elevation in graph.

As we can’t conclude the index values properly, this graph again simplified for single state as follows:

From these graph, we can conclude the index values of different sectors in each state separately. We can draw further states.

s

***Bibliography***

<https://data.gov.in/resources/state-level-consumer-price-index-ruralurban-december-2017-base-2012100>

<https://quickonomics.com/calculating-consumer-price-index-cpi/>

<http://www.arthapedia.in/index.php?title=Consumer_Price_Index(Urban)_and_Consumer_Price_Index(Rural)>

<https://exceljet.net>

<http://www.etldatabase.com/etl-load/>

<http://www.etldatabase.com/etl-transform/>

<http://www.etldatabase.com/etl-extract/>

<http://www.etldatabase.com/etl-process/>

<https://www.investopedia.com/terms/i/inflation.asp>

<https://en.wikipedia.com/wiki/Data.gov.in>