World Population Analysis

Introduction

The third project done by me using Power BI tool is the World Population Analysis. This project contains huge amount of data. The project is been analysed thoroughly using the Power BI tool for better analysis and understanding of data.

Objective

*The main objective of doing this project of World Population Analysis is to know the population of each countries and to determine particularly the population of male and female.* Now, we’ll get into the whole process involved in the World Population Analysis.

* Extract: First we’ll open the Power BI Desktop. In the Power BI Desktop of Home Tab click on Get Data option for extracting the data from different data sources. So we’ll click on Excel option for extracting data from excel files like Population 1950-1999, Population 2000-2049 and Population 2050-2100 and a text file called code-country-region.
* Transform: After extracting the data one by one from the excel spreadsheet, we’ll load it into the Power Query Editor by clicking on transform option for the data preparation and data transformation. Now, we can rearrange the data in the Power Query Editor as per the requirements.
* Load: After completion of the data preparation in the Power Query Editor we’ll load the prepared data in the Power BI Desktop by clicking on Close and Apply option in the Home Tab.

Here are the steps or process involved in the preparation of data:

We have three files of population data like Population 1950-1999, Population 2000-2049 and Population 2050-2100. In all the three files we had small changes to be made like “Use first row as headers”, “Remove null values”, “Replace values” etc. Each data population had approximately 2 lakh rows. After completing all the above tasks we used an option called Append Queries. Append Queries helps the user in appending one file with the other for reducing the workload of the employees. To append, the queries must have same column name and data types in all the other files. This appended query is called as “Combined Population”. In the Combined Population we approximately have more than 12lakh rows.

1. Remove Columns: The first step after appending is to remove the unwanted columns which doesn’t have any use with the data in the future.
2. Unpivot Columns: The second step is to unpivot the columns of Population male, Population female and Population Total into Attribute and Value Pairs. We removed the Population total column as the sum of Population of male and female is the population.
3. Split Column by Position: The third step is to split the column of attribute from Popmale and Popfemale to Male and Female. The split column by position helps in removing the no. of characters we assign.
4. Rename Columns: The fourth step is to rename the column of attribute to Gender and value column to Population.
5. Replaced Value: The fifth step is to replace the values of the age group column from 14-Oct to 10-14 and 9-May to 5-9.

Next we’ll create two new groups called Data sources and Data Models. The data sources contains all the three population data and the combined population. The data models contains the Dimensional Age, Dimensional Region and Fact Population. The data model is prepared using a technique called Star Schema. We have another group called other groups which contains the files like code-country-region and region fullname.

Star Schema is one of the most important topic in Power BI. It is a technique which is used in data warehouses or business intelligence. It is a database that uses one large fact table to store measured data and one or more small dimension tables that stores attributes about the data and which are connected to the fact table. This is known as star schema.

The first data model we are going to talk about is the Dimensional Region. The steps involved in preparing it are as follows:

* Extract: We extracted the data from the data sources i.e., combined population. We took a reference of the combined population as this is our original data and any changes in the combined population will be changed in the other data also.
* Remove Columns: We removed all the columns except country id and country as we wanted the data related only to the region.
* Merge Queries: After removing all the unwanted columns we wanted a column called region in the data. So, we used an option called Merge Queries. This merge query helps in connecting two or more tables or files with one same column in both the files. We extracted the data called code-country-region in the other sources option. In code-country-region we had a column called region. We merged dim region with code-country-region with having the country id as the same column. The merged column will be shown in the dim region data or file. Here we used left kind join as it says keeping the left query same and common rows of right query.
* Expanded code-country-region: After completing the merging step we should expand the merged column to region by selecting in the filter option of the column.
* Remove null values or Filter Rows: This tells us to remove the null or n/a values present in the merged column as it doesn’t have any value or make any sense.
* Remove Duplicates: This step helps us in removing the duplicates present in the data. It removes all the repeated values present in the data and gives only the unique or a single value.

The second data model we are going to talk about is the Dimensional Age. The steps involved in preparing it are as follows:

* Extract: We extracted the data from the data sources i.e., combined population. We took a reference of the combined population as this is our original data and any changes in the combined population will be changed in the other data also.
* Remove Columns: We removed all the columns expect age group as all other columns were not needed and which were not related to the age.
* Remove Duplicates: We removed the duplicates i.e., the repeating values and keep only the unique values present in the age group column.
* Add Index: We used a new option called add index present in the add column tab. In the add index we should select from 1 to add a new column.
* Rename Column: We renamed the new added column as Age ID as it contains unique ID’s.
* Reorder Column: This option helps the user to reorder the columns as per the requirements. We interchanged the age group and age id column.
* Inserted text after delimiter: We used this option in the age group column as we needed only the ending value of ages i.e., 4,9,14,19 etc., so we used text after delimiter option because we had “hyphen” inbetween the ages.
* Replace Value: This option helps the user to replace one value from another. We have replaced the null value present in the last row of text after delimiter column to 100.
* Add Conditional Column: In the Add Column tab, we should choose conditional column to give condition according to the requirement. After clicking on conditional column we should give a new column name and give condition based on other column or values.
* Remove Column: This option helps the user to remove the unwanted column which doesn’t have any use. We removed the text after delimiter column as it is not used in the further process.

The third data model we are going to talk about is the Fact Population. The steps involved in preparing it are as follows:

* Extract: We extracted the data from the data sources i.e., combined population. We took a reference of the combined population as this is our original data and any changes in the combined population will be changed in the other data also.
* Remove Column: It helps in removing the unwanted columns present in the data. We removed the country column as we had country id for each countries.
* Merged Queries: This concept helps in merging two or more queries having one same column in both the data. We merged age id from the dim age data to the fact population data having age group as the same column.
* Expanded Dim Age: After merging a new column into the fact population, we should expand the new column by clicking on the filter option in the column and selecting only age id.
* Remove Column: Again we removed the age group column as we had age id for each age group.
* Reorder Column: After completing all the above steps we reordered the columns in a specific way as per the requirements.

This is all about the steps involved in preparing the data for the purpose of visualization. After the preparation of data we’ll close and apply to the Power BI Desktop. In the Power BI Desktop we should go through the data model to see whether the cardinality relationships are correct or not. We saw that the relationship between the tables were matching and having many to one relationships. Fact population is the factorial table and dim age, dim region are the dimensional tables. After looking into the data model we’ll move on to the Report View for creating different types of visuals.

Data Visualization

We have used many different types of visuals for creating a interactive and meaningful dashboard. The whole dashboard should give us a clear idea about what is been done.

* The first visual used in the dashboard is the text visual used for writing heading of the data.
* The second visual is the line chart used for plotting population by age group. This chart tells us the increasing or decreasing order in the form of line.
* The third visual is the clustered bar chart used for plotting top 10 countries with highest population. This chart shows the differences of population between the countries in the cluster form.
* The forth visual is the pie chart used for plotting population by age group. Pie chart is used only when we have less than seven categories as it helps in showing the differences of a part of a whole. In our data we have only six categories of age group and pie chart is used for plotting it.
* The fifth visual is the clustered column chart used for plotting population by region. The size of the column indicates the number of population in the particular region.
* The sixth visual is the map which shows the population of each country by the size of the bubbles.
* The next type of visual used in the dashboard is the slicers. We have used slicers for the country and years as to know the population of each country in a particular year.
* We have used card visual for depicting the sum of population, male population and female population.
* For knowing exactly the male and female population we have used a concept called DAX. DAX is a Data Analysis Expression which contains set of functions and formulas used for calculation. By the use of DAX it is very easy for a user to compute the task very fast and quickly.

This is all about a brief description on data preparation, data cleaning, data transformation and data analysis for creating interactive dashboard. Now it is very easy for any user to know the population of each country by just one click with the help of Power BI.