2.4.1 -Cleaning data

During the cleaning exercise, we hope to turn the available data into a dataset that is usable.

This means that during this stage of the crisp dm we need to remove what we do not want from

the dataset and decide on what we want to keep

**ML use for CA1**

**Predictive Modeling:**

Useful when you want to **make predictions** or forecasts based **on historical population data**.

This could involve **predicting future population growth**, healthcare demand, or the impact of policy changes on the population.

**Predicting future population trends** or healthcare needs **based on historical demographic** and health data. **We can use it for both datasets.**

**Anomaly Detection:**

Anomaly detection can help **identify unusual or unexpected patterns** or outliers **in population data**.

It's useful for **detecting rare events** **or deviations from typical population behavior.**

**May use for the emigration around covid!**

**Supervised learning algorithm - best choice in our case.**

**For predicting future population growth based on historical population data**, supervised learning algorithms are **the most appropriate choice**.

Supervised learning is a type of machine learning where the algorithm is trained on a labeled dataset, meaning that it learns from historical data with known outcomes (in this case, historical population data and corresponding population growth data) to make predictions about future outcomes.

The Plan:

**Anchor points:**

1. **Data Preparation:**

Collect historical population data, which includes attributes such as year, location, demographic factors.

Label the data with the corresponding population growth values for each year or period.

***\* Load and preprocess the dataset. Ensure that the data is clean, and handle any missing values if present.***

1. **Data Exploration:**

*\* Analyze the dataset to gain insights into population trends and patterns over the years.*

1. **Feature Engineering:**

Identify relevant features (input variables) that can influence population growth.

These features might include birth rates, death rates, migration rates, economic indicators, healthcare data, etc.

Perform any necessary data preprocessing, such as **handling missing values**, **scaling features**, and **encoding categorical variables**.

\* **Create any additional features that might help your model capture population growth trends more effectively.**

For example, you could calculate population growth rates.

1. **Split the Data:**

Split the historical dataset into **training data** and **testing data** to evaluate the model's performance.

\* **Split dataset into training and testing sets. The training set will be used to train the machine learning model, and the testing set will be used to evaluate its performance.**

1. **Algorithm Selection:**

Choose a regression algorithm from the supervised learning category.

Regression algorithms are specifically designed for predicting continuous numerical values, making them suitable for population growth prediction.

**\* Common regression algorithm:**

**Linear regression** is a common choice for population growth prediction.

We can also explore more complex algorithms like **decision trees**, **random forests**, or **neural networks.**

1. **Model Training:**

Train the selected regression model using the training dataset.

The model will learn to make predictions based on the historical population data.

**\* Train chosen regression model on the training data, using the years as input features and population counts as the target variable.**

1. **Prediction**

Once the model has been trained and evaluated, you can use it to make predictions about future population growth by providing it with relevant input features.

**\* Use the trained model to make predictions for future years beyond the range of your dataset (e.g., 2050).**

These predictions will represent the estimated population counts for this year.

1. **Model Evaluation:**

Use the testing dataset to assess the model's performance.

Common evaluation metrics for regression tasks include **Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2).**

These metrics will help to understand how well the model predicts future population growth.

1. **Visualization:**

**Create visualizations of the model's predictions**, historical population data, and trends. This can help to communicate the results effectively.

1. **Fine-Tuning:**

Depending on the model's performance, we may need to fine-tune hyperparameters, using different algorithms, or experiment with more advanced regression techniques.

1. **Prediction Scenarios: ???**

Consider running different scenarios based on various assumptions or external factors that may affect future population growth. For example, we can simulate the impact of different birth rates, immigration rates, or policy changes on population projections?

1. **Communicate Results:**

Present the population growth predictions along with insights and potential implications.