# **LifeSure Project: Phase 1 - Data Collection and Integration**

Written by EL AYOUBI Hilal, BROSSIER Léandre, BOURDET Caroline, COCURON Morgane, FIGARD Clément, SKANDERI Elyes

## **1. Introduction**

This document outlines the roadmap for the first phase of the LifeSure project, focusing on **data collection and integration**. The goal of this phase is to gather, clean, and integrate relevant external data sources to provide a comprehensive understanding of customer behaviors, market trends, and sustainability factors. This will serve as the foundation for the subsequent phases of data analysis, visualization, and dashboard creation.

## **2. Objectives**

* **Collect** diverse external datasets relevant to LifeSure's needs, including:
  + Market trends
  + Customer demographics
  + Environmental data
  + Public sentiment
* **Clean and prepare** the data to ensure it is ready for analysis.
* **Integrate** the data into a unified dataset suitable for visualization tools.

## **3. Data Sources**

The following datasets will be collected from **Kaggle** and other public sources:

### **3.1 Customer Demographics and Behavior**

* **Dataset**: [customer personality](https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis)
  + **Description**: Contains data on customer demographics, purchasing behavior, and preferences.
  + **Use Case**: Segment customers based on their behavior and preferences.

### **3.2 Market Trends and Insurance Industry Data**

* **Dataset**: [travel insurance trend](https://www.kaggle.com/datasets/tejashvi14/travel-insurance-prediction-data)
  + **Description**: Contains data on travel insurance, customer demographics, and risk factors.
  + **Use Case**: Analyze trends in insurance pricing and customer risk profiles in the travel sector.
* **Dataset**: [vehicle insurance](https://www.kaggle.com/datasets/imtkaggleteam/vehicle-insurance-data)
  + **Description**: Contains data on vehicle insurances, customer demographics, and risk factors.
  + **Use Case**: Analyze trends in insurance pricing and customer risk profiles in the car sector.
* **Dataset:** [health insurance](https://www.kaggle.com/datasets/adsamardeep/insurance-and-medical-cost)
  + **Description**: Contains data on health insurance, customer demographics, and risk factors.
  + **Use Case**: Analyze trends in insurance pricing and customer risk profiles in the medical sector.

### **3.3 Environmental Data**

* **Dataset**: [Global Environmental Indicators](https://www.kaggle.com/datasets/theworldbank/global-environmental-indicators)
  + **Description**: Contains data on various environmental indicators, such as CO2 emissions, energy use, and forest area.
  + **Use Case**: Study environmental factors that could influence insurance policies.

### **3.4 Public Sentiment**

* **Dataset**: [Twitter US Airline Sentiment](https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment)
  + **Description**: Contains tweets about US airlines, with sentiment labels (positive, negative, neutral).
  + **Use Case**: Analyze public sentiment about insurance companies or sustainability topics.

## **4. Tools and Technologies**

The following tools will be used for data collection, cleaning, and integration:

* **Python Libraries**:
  + **Pandas**: For data manipulation and cleaning.
  + **NumPy**: For numerical operations.
  + **Matplotlib/Seaborn**: For initial data visualization.
* **Excel**: For manual data cleaning and quick exploration.
* **Kaggle**: For downloading datasets.

## **5. Data Collection and Integration Plan**

### **5.1 Data Collection**

* Download datasets from Kaggle and other public sources.

### **5.2 Data Cleaning**

* Handle missing values:
  + Fill in missing data using appropriate methods (e.g., mean, median, or mode).
  + Remove incomplete records if necessary.
* Correct inconsistencies:
  + Standardize date formats, currency, and units.
  + Fix typos and errors in the data.
* Remove duplicates and irrelevant data.

### **5.3 Data Integration**

* Combine datasets from different sources into a unified format.
* Create new variables to enhance analysis:
  + **Customer segments**: Group customers based on demographics or behavior.
  + **Environmental impact scores**: Create a score based on environmental data.

## **6. Next Steps**

Once the data collection and integration phase is complete, the next steps will include:

* **Phase 2: Data Analysis and Specifications**:
  + Perform exploratory data analysis (EDA) to uncover trends and patterns.
  + Create a specifications document for the dashboard.
* **Phase 3: Visualization and Communication of Results**:
  + Develop an interactive dashboard using tools like Tableau or Power BI.
  + Present key findings and recommendations to stakeholders.

## **7. Conclusion**

We started exploring each dataset to check it would fit our needs for the project. This roadmap serves as a steppingstone to check all is in order for the start of our project with data cleaning. All data will follow the same schema to verify it is correctly cleaned as depicted in 5.2.

## **Analysis-Phase 1**

### **5.2 Data Cleaning**

We used Google Colab and Jupyter Notebook to perform the data preparation and cleaning.

GoogleColab link : https://colab.research.google.com/drive/1KSDrrlhTng0jCyGdNPqwh4IwsE0HprwX?usp=sharing

1. 1rst dataset: <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>.