**Project 3-Data Engineering Track**

**PROPOSAL-GLOBAL INFLATION**

1. **Tasks by Team Member:**
   * Jupyter Notebook data cleaning: Liseth & Kim
   * ERD & SQL Schemata: Derilee
   * Jupyter Notebook data visualization: Love
   * SQL Queries: Liseth
   * Research data: Sacide
   * PowerPoint presentation: Everyone
2. **Number of unique records:**
   * Country continent dataset contains 249 unique records.
   * Global inflation dataset contains 206 unique records.
3. **New library used:** 
   * Seaborn
4. **Method for reading data from the database:**

* Pandas DataFrame

1. **Project overview:**

The project aims to explore global inflation, defined as the increase in prices over time, which can impact significantly both economic systems and individuals. Utilizing data from two CSV databases, the analysis employs Python, Jupyter Notebook, and SQL to conduct data cleaning, management, and visualization. The findings provide valuable insights into various inflationary trends on a global scale of different inflation types using descriptive statistics and data visualization. Additionally, the project includes predictive analytics on the country with the highest inflation rate, thereby providing valuable foresight into future economic conditions.

1. **Project purpose:**

The purpose of the project is to analyze global inflation trends to understand how rising prices affect economies and individuals. The project aims to:

* Identify and analyze different types of inflation across the world.
* Generate insights and provide answers to:
  + Which countries have the highest and lowest inflation rates from 2002 to 2022.
  + What are the average inflation rates across the world from 2002 to 2022.
  + Which countries have the highest and lowest inflation rates during COVID-19
  + What are the future inflation rate predictions on the country with the highest inflation rate.

1. **Project Questions:**
   1. What is the average, median, maximum, minimum, range, standard deviation, and variation from 2002 to 2022 for the following five types of inflation:
      1. Energy Consumer Price Inflation
      2. Food Consumer Price Inflation
      3. Headline Consumer Price Inflation
      4. Official Core Consumer Price Inflation
      5. Producer Price Inflation
   2. Which country has the highest inflation rate from 2002 to 2022 for each of the five types of inflation in question A?
   3. What is the average inflation rate during COVID-19 (2019-2022) for the following types of inflation:
      1. Energy Consumer Price Inflation
      2. Food Consumer Price Inflation
      3. Headline Consumer Price Inflation
   4. Which country has the highest inflation rate during COVID-19 for each of the three types of inflation in question C?
   5. Which country has the lowest inflation rate during COVID-19 for each of the three types of inflation in question C?
   6. Which country saw the highest inflation increase during COVID-19 overall?
   7. What are the predictive statistics on the country with the highest inflation rate in COVID-19 from question F?
2. **Project instructions:**

* **Prerequisites:** 
  + Python
  + Jupyter Notebook
  + SQL
  + **Required Libraries:** pandas, seaborn, matplotlib, scikit-learn
  + **Required Modules:** pathlib, matplotlib.pyplot, matplotlib.ticker, sklearn.linear\_model.
* **Folders:**
  + **Resources\_csv\_files:** 
    - Original CSV files:
      * Country\_continent\_original\_csv
      * Global\_dataset\_of\_inflation\_original\_csv
    - New CSV files:
      * Country\_codes
      * Global\_inflation\_by\_continent
      * Energy\_consumer\_price\_inflation
      * Food\_consumer\_price\_inflation
      * Headline\_consumer\_price\_inflation
      * Official\_core\_consumer\_price\_inflation
      * Producer\_price\_inflation
  + **SQL\_files:**
    - **Schemata:** SQL tables
    - **Queries:** Data analysis
* **Additional Files:**
  + ERD\_image:
    - Entity-Relationship Diagram to support ETL workflow
  + Jupyter\_Notebook\_file:
    - Data cleaning and data visualization
  + Project\_3\_PowerPoint
  + Project\_3\_proposal

1. **Database choice:**

* Using SQL for the global inflation rates project offers several key benefits:
  + For complex queries, SQL enables advanced data operations, such as aggregations and descriptive statistical calculations, which are crucial for in-depth analysis.
  + For storage purposes, SQL databases organize data in a structured manner, facilitating efficient management and retrieval of information.
  + For advanced queries, SQL supports sophisticated data analysis, including filtering, grouping, and trend identification, essential for accurate and comprehensive insights.
* By utilizing SQL, data management and analysis ensures reliable and insightful results for the project.

1. **Ethical considerations:**

When using SQL and Jupyter Notebook with the two datasets from Kaggle, the ethical considerations for the project included ensuring data privacy and security practices, such as protecting personal information and managing data securely. We verified the accuracy and credibility of the datasets by thoroughly assessing the sources and ensuring that the original CSV files remained unmodified. Additionally, transparency in documenting our methods and code of our findings were also crucial to maintaining ethical standards throughout the project.

1. **References for the data source:** 
   * [Global Inflation Dataset - (1970~2022) (kaggle.com)](https://www.kaggle.com/datasets/belayethossainds/global-inflation-dataset-212-country-19702022) – Author: Belayet Hossain
   * [country to continent (kaggle.com)](https://www.kaggle.com/datasets/statchaitya/country-to-continent)
2. **References for external code:**
   * [**https://seaborn.pydata.org/tutorial.html**](https://seaborn.pydata.org/tutorial.html)
3. **ETL workflow documentation with ERD Image:**

A screenshot of a computer

Description automatically generated

Can you describe your experience with troubleshooting technical issues?

For a surgical case, the anesthesia provider called me to troubleshoot the Drager Perseus A500 machine. When I got to the room, the problem they were facing was fluctuations with the CO2 levels and the patient slowly turning bluish.

With this information, instead of going through the whole system check for entire machine, there are three important things that I always considered when troubleshooting:

1. The physical appearance of the patient, meaning skin color, temperature, level of consciousness.
2. Screen, showing vital signs (heart rate, blood pressure, respiratory rate, end-tidal CO2).
3. Analyzing how the machine’s performance and operation can affect these conditions.

Therefore, I remember the patient was first turning bluish and the CO2 levels were increasing exponentially.

* If the CO2 levels are too low:
  + The patient may turn slightly bluish
  + The component of the machine that may affect this could be the oxygen cylinder being too open and therefore increasing the concentration of oxygen going into the patient through the respiratory system, which would essentially dilute the amount of CO2 and therefore decrease those CO2 levels.
* If the CO2 levels are too high:
  + The patient may turn bluish and purple.
  + The component of the machine that could affect this would be the CO2 absorber. When a patient is under anesthesia, CO2 is typically exhaled by the patient and enters the anesthesia machine through the respiratory system. The CO2 absorber then captures and removes CO2 from the gas mixture before it is recirculated back to the patient. Therefore, if the CO2 absorber is expired, meaning too saturated with CO2 turning blue or purple, the concentration of CO2 exhaled by the patient would be recycled back to the patient, increasing those CO2 levels on the screen.