Data-Driven Analysis of Employment Conditions and Food Security in the United States (2022 Analysis)

Research Question

How do employment conditions affect household food security in the United States in 2022? This study aims to investigate the relationship between employment characteristics, medical conditions, and demographic factors and their impact on food security status in the United States, using data from 2022. Understanding these relationships is crucial for developing effective policies and interventions to improve food security.

Data Sources

1. **Job and Employment Conditions Dataset (HC-237, 2022)**
    - **Metadata URL**: Link to dataset metadata
    - **Data URL**: Link to dataset
    - **Description**: This dataset contains detailed information on employment status, wages, job types, employer-provided benefits, and other employment-related factors for individuals in the United States in 2022. It helps analyze trends in job stability, wages, and benefits, and their impact on household food security.
    - **Structure**: Tabular format with rows representing individual respondents and columns representing various employment metrics.
2. **Food Security Data (HC-240, 2022)**
    - **Metadata URL**: Link to dataset metadata
    - **Data URL**: Link to dataset
    - **Description**: This dataset includes household-level information related to food security in 2022, such as worries about food running out, skipping meals, and affordability of balanced meals. It helps analyze the prevalence of food insecurity across different employment conditions.
    - **Structure and Quality**: The dataset is structured in tabular format, with columns providing detailed food security indicators.

Reasons for Choosing These Data Sources

- **Relevance**: Both datasets are from the **Medical Expenditure Panel Survey (MEPS)**, making them highly relevant for analyzing socioeconomic conditions, employment, and food security.
- **Coverage Period**: The data spans the same time frame (2022), which allows for consistent temporal analysis.
- **Open Data**: Both datasets are publicly available under an open-data license, provided by a reputable source, ensuring transparency and accessibility.

Data Quality and Licenses

The datasets from MEPS are provided under an **open-data license**, meaning they can be used freely as long as proper attribution is given. According to the **MEPS Copyright Notice** (source), the Agency for Healthcare Research and Quality (AHRQ) has stated that all published material is in the public domain except for previously copyrighted photographs and illustrations. This public domain material is free for use without specific permission, although it is requested that users cite AHRQ and the Medical Expenditure Panel Survey as the source. To comply with these license requirements, all publications and reports generated from this data will clearly attribute MEPS and provide links to the original data.

The datasets are structured, clean, and consistent, ensuring minimal pre-processing needs, although missing values are still possible. For additional accessibility information, refer to the **MEPS Accessibility Notice** (source).

Data Pipeline

The data pipeline was implemented in **Python**, utilizing libraries like **Pandas** for data manipulation, **Requests** for data download, and **SQLite** for data storage.

The data pipeline consisted of the following major steps:

1. **Data Acquisition**: The datasets were downloaded from their respective URLs as **ZIP files** containing **Excel sheets**. Using the requests library, the ZIP files were accessed, and the relevant Excel sheets were extracted.
2. **Transformation and Cleaning**:
   - The job dataset and the food security dataset were cleaned to remove unnecessary columns and rename them for clarity.
   - **Missing Values**: Columns containing numeric information, such as hours worked and meal skipping frequency, were converted to numeric types while handling missing values using errors='coerce' and subsequently dropping rows with missing data.
   - **Standardization**: Columns were standardized to ensure consistency between the two datasets before merging.
3. **Merging**: The two datasets were merged on the dwelling_unit_id field, representing the common identifier between the employment and food security data.

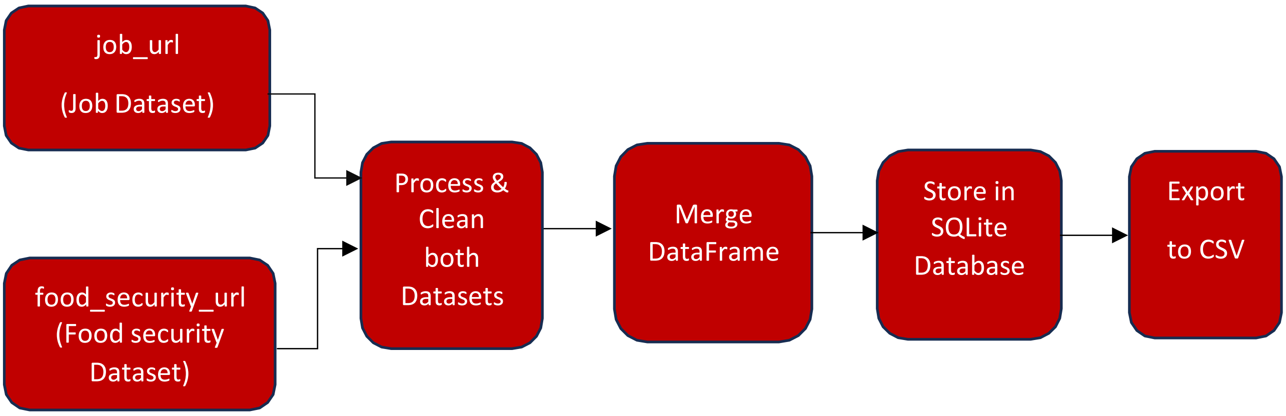
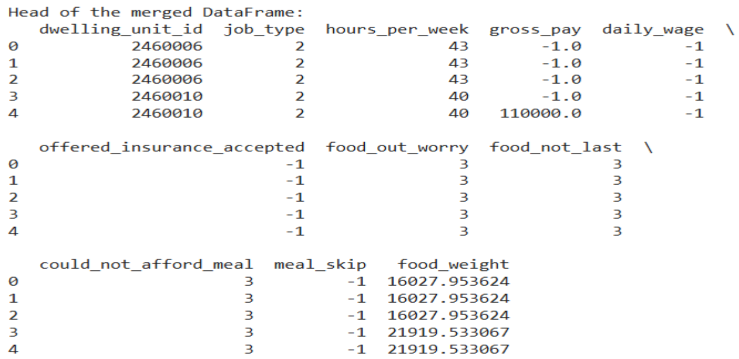


Fig: Automated Data Pipeline

4. **Storage**: The cleaned and merged dataset was stored in both **SQLite** (merged_dataset.db) for easy querying and in **CSV format** (merged_dataset.csv) for further analysis and visualization.



```
Head of the merged DataFrame:
   dwelling_unit_id  job_type  hours_per_week  gross_pay  daily_wage  \
0           2460006         2              43       -1.0          -1
1           2460006         2              43       -1.0          -1
2           2460006         2              43       -1.0          -1
3           2460010         2              40       -1.0          -1
4           2460010         2              40   110000.0          -1

   offered_insurance_accepted  food_out_worry  food_not_last  \
0                          -1               3              3
1                          -1               3              3
2                          -1               3              3
3                          -1               3              3
4                          -1               3              3

   could_not_afford_meal  meal_skip   food_weight
0                      3         -1  16027.953624
1                      3         -1  16027.953624
2                      3         -1  16027.953624
3                      3         -1  21919.533067
4                      3         -1  21919.533067
```

Fig: Merged Dataset

Technology Used

- **Python**: Data processing and transformation.
- **SQLite**: Storing the final merged dataset for efficient access and queries.
- **Pandas and Requests Libraries**: For data cleaning, manipulation, and retrieval.

Problems Encountered and Solutions

- **Missing Values**: Both datasets had missing values that needed to be handled. This was resolved by converting columns to numeric and dropping rows with NaN values.
- **Data Merging Issues**: Initially, discrepancies in column names and data formats created problems during merging. This was resolved by renaming columns and ensuring consistency across both datasets.

Meta-Quality Measures

The pipeline includes several measures to ensure data quality and adaptability:

- **Data Validation**: Columns were validated to confirm they match expected data types after loading and transforming.
- **Error Handling**: The pipeline uses exception handling to manage data retrieval failures or unexpected errors in data processing, allowing the process to log errors instead of crashing.
- **Adaptability**: The pipeline can adapt to minor schema changes (e.g., additional columns), ensuring robustness in case of minor modifications to the input datasets.

Results and Limitations

- **Output Data**: The output of the data pipeline consists of a cleaned, merged dataset saved in SQLite and CSV formats. The dataset contains fields for both employment and food security metrics, such as hours worked, gross pay, meal skipping frequency, and food weight.
- **Data Quality**: The final dataset maintains high-quality standards: it is accurate, complete, and consistent. However, there are limitations, such as potential biases in the original data collection and the lack of real-time updates.
- **Data Format**: The output is provided in **CSV** format for interoperability and **SQLite** for efficient querying. These formats were chosen because they balance usability for analysis and storage efficiency.