**Summary Report of Region Clustering and Agricultural Analysis, and insights obtained from this**

**Report 1: Agricultural Production Clustering and Region Analysis**

**1. Objective**

The goal of this project was to preprocess and cluster agricultural production data, analyze regional production capacities, and visualize clustering results for informed decision-making.

**2. Data Loading and Initial Exploration**

* **Primary Data**: merged\_data.csv  
  The main dataset containing agricultural production records, environmental variables, and other features.
* **Supplementary Data**: region\_dim.csv  
  This dataset provided Region\_ID and Region\_Name, used to map production data to specific regions.

**Initial Observations**:

* Data loaded successfully with pd.read\_csv.
* Printed the first (df.head()) and last (df.tail()) records to inspect the structure and content.

**3. Data Preprocessing**

**Handling Missing Values**:

* Rows with missing values were dropped using df.dropna().  
  Alternative strategies like imputation could be explored depending on the data context.

**Region Name Processing**:

* Non-string entries in Region\_Name were removed.
* Duplicates were dropped to ensure clean region mappings.
* Fuzzy matching (fuzzywuzzy.process) was applied to reconcile mismatches between Region\_Name in the main dataset and supplementary dataset.

**4. Data Transformation**

**Categorical Encoding**:

* Used OneHotEncoder to transform categorical variables into numerical representations.

**Feature Selection**:

* Removed features with low variance (VarianceThreshold) to retain only impactful variables.

**Standardization**:

* Applied StandardScaler to normalize feature values for improved clustering performance.

**Dimensionality Reduction**:

* Performed PCA to reduce dimensions while retaining 95% of the variance.

**5. Clustering Analysis**

**Metrics for Evaluation**:

* **Silhouette Score**: Measures how well clusters are separated.
* **Davies-Bouldin Index**: Lower values indicate better clustering.
* **Calinski-Harabasz Index**: Higher values indicate better-defined clusters.

**Cluster Range (k=2 to 10)**:

* Metrics were computed for each k, and results were plotted to identify the optimal number of clusters.

**Optimal Clusters**:

* Based on the evaluation, **k=4** was selected as the optimal number of clusters.

**6. Cluster Labeling and Visualization**

* Clusters were labeled with meaningful names:
  + **Cluster 0**: High Capacity
  + **Cluster 1**: Moderate Capacity
  + **Cluster 2**: Low Capacity
  + **Cluster 3**: Emerging
* **t-SNE Visualization**:
  + A two-dimensional scatterplot was generated to visualize clusters using t-SNE.
  + The plot displayed distinct separation among clusters, confirming their uniqueness.

**7. Region Matching and Unmatched Data**

* **Region Mapping**:
  + The main dataset was merged with region\_dim.csv to associate clusters with specific regions.
* **Unmatched Regions**:
  + Regions in the supplementary dataset but not matched to the main dataset were identified and saved in unmatched\_regions.csv.

**8. Outputs**

1. **Clustered Data**:
   * Saved to data\_with\_clusters.csv.
   * Includes features, clusters, and region information.
2. **Unmatched Regions**:
   * Saved to unmatched\_regions.csv.
   * Highlights regions requiring further data reconciliation.

**9. Visual Insights**

1. **Cluster Validation Metrics**:
   * A line plot showed the trends of evaluation metrics (Silhouette, Davies-Bouldin, Calinski-Harabasz) across varying k.
   * The elbow point was observed at k=4.
2. **t-SNE Cluster Visualization**:
   * Clusters were clearly defined, indicating successful clustering.
   * The plot grouped similar regions into distinct clusters based on production capacity.

**10. Recommendations and Future Work**

* **Data Imputation**: Explore imputation methods to handle missing data rather than outright removal.
* **Further Analysis**:
  + Investigate the relationship between environmental factors and cluster membership.
  + Assess how external factors (e.g., market access) influence cluster performance.
* **Scalability**: Test the model with larger datasets or new regions to validate robustness.
* **Refinement of Matching**:
  + Apply more advanced fuzzy matching techniques for unmatched regions.
  + Validate unmatched regions with domain experts.

This analysis successfully clustered regions based on production capacity and highlighted distinct regional profiles. The results provide actionable insights for regional planning, resource allocation, and policy formulation in agricultural production.

**Detailed Report 2: Insights from Analyzing Agricultural Clusters**

This report summarizes insights derived from running SQL queries on the CSV file data\_with\_clusters.csv, which was generated from the previous clustering process.

**1. Regions with Highest Production Variability**

**Objective**

To identify regions that experience the most significant variability in agricultural production.

**Query**

The query calculated the maximum production variability (Production\_Variability) for each region.

**Results**

* The table ranks regions by Max\_Production\_Variability, allowing identification of regions most affected by inconsistent production levels.
* These insights are essential for targeting stability-focused interventions or policy support.

**Key Takeaways**

Regions with the highest variability might require:

* Better environmental or resource management practices.
* Improved infrastructure to mitigate factors contributing to volatility.

**2. Relationship Between Temperature and Production**

**Objective**

To examine how the average temperature impacts total agricultural production in each region.

**Query**

The query aggregated:

* Avg\_Temperature for each region.
* Total production (Total\_Annual\_Production).

**Results**

* Regions were sorted by Avg\_Temperature, and their total production was calculated.
* Trends could show how temperature levels influence productivity.

**Key Takeaways**

* Regions with moderate temperatures likely have optimal conditions for high productivity.
* High or low extremes might correlate with reduced production, hinting at thermal stress on crops.

**3. Correlation Between Rainfall and Production**

**Objective**

To analyze the effect of rainfall levels on production.

**Query**

The query grouped data by Annual\_Rainfall and computed the average production (Avg\_Production).

**Results**

* Rainfall levels were sorted alongside average production.
* The results may indicate a threshold of rainfall optimal for maximizing production.

**Key Takeaways**

* Insufficient rainfall corresponds to reduced production.
* Excess rainfall may also negatively impact production, implying the need for balanced irrigation systems.

**4. Cluster Membership Analysis**

**Objective**

To examine the distribution of regions across different production clusters.

**Query**

The query counted the number of unique regions within each cluster (Cluster\_Label).

**Results**

* Clusters were ranked by the number of associated regions, offering insight into the regional distribution of production capacity.

**Key Takeaways**

* Larger clusters might indicate widespread similar conditions, whereas smaller clusters suggest specialized conditions.
* This distribution provides a roadmap for targeted regional development.

**5. Highest Production Region by Cluster**

**Objective**

To identify the region with the highest production within each cluster.

**Query**

The query calculated the maximum Total\_Annual\_Production for each region within each cluster.

**Results**

* This provides a leaderboard of top-performing regions in each cluster.
* Helps identify best practices or conditions contributing to high productivity.

**Key Takeaways**

* High-performing regions could serve as benchmarks or examples for others in the same cluster.
* Policies should support regions with lower performance to narrow the productivity gap within clusters.

**6. Extreme Weather and Its Impact**

**Objective**

To understand how extreme temperature variations affect production.

**Query**

The query calculated:

* Minimum and maximum temperatures for each region.
* Average production (Avg\_Production) for these regions.

**Results**

* Regions experiencing extreme temperature fluctuations were identified.
* The relationship between temperature range and average production was analyzed.

**Key Takeaways**

* Regions with extreme temperatures may require climate-resilient crops or adaptive agricultural practices.
* Strategies to mitigate the effects of extreme weather could stabilize productivity.

**General Observations**

1. **Data Quality**:
   * The dataset used was well-processed, with clean Region\_Name associations from the previous script.
   * Results demonstrate the importance of preprocessing in generating actionable insights.
2. **Cluster Impact**:
   * Clustering enabled focused analysis, grouping regions by shared characteristics.
   * Each cluster's distinct properties allow tailored strategies.
3. **Environmental Factors**:
   * Both temperature and rainfall emerged as critical factors influencing production.
   * These variables should be a focal point in agricultural planning.
4. **Actionable Insights**:
   * High variability and extreme weather regions are key areas for intervention.
   * Top-performing regions could provide strategies to uplift struggling areas.

**Outputs**

* **Analysis Results**: The results were displayed in tabular format during execution. For permanent storage:
  + Convert result tables to CSV for documentation.
  + Use these for further reporting or visualization.
* **Database Setup**: The use of an SQLite in-memory database allowed efficient querying and testing without needing a persistent setup.

**Recommendations**

1. **Data Utilization**:
   * Integrate results into dashboards for visual analysis (e.g., using tools like Tableau, Power BI).
   * Conduct year-over-year comparisons to track changes in production.
2. **Further Analysis**:
   * Examine other environmental and socioeconomic factors influencing clusters.
   * Conduct predictive modeling for future production trends.
3. **Targeted Interventions**:
   * Invest in irrigation and climate-resilient crops for extreme weather regions.
   * Expand support for high-variability regions to stabilize productivity.

This detailed report encapsulates the value of structured querying and analysis for data-driven decision-making in agricultural contexts.