Real-Time Data Quality Assessment and Query Interface

1. Context & Overview

This project implements an intelligent data cleaning and query system that combines automated data quality assessment with natural language processing capabilities. Built for data processing needs, the system aims to streamline data analysis workflows by automating anomaly detection and enabling conversational data queries.

Key Objectives:

- Automate data cleaning and anomaly detection processes
- Provide an intuitive interface for data analysis
- Enable natural language querying of datasets
- Ensure data quality and consistency

2. Technical Implementation

2.1 System Architecture

The solution is built using modern technologies and follows a modular design:

1. **Frontend Layer**
   - Streamlit-based web interface
   - Responsive design with custom styling
   - Interactive file upload and query components
2. **Processing Layer**
   - Data cleaning and validation pipeline
   - Real-time anomaly detection
   - Natural language query processing
3. **AI Integration Layer**
   - Azure OpenAI service integration
   - LangChain for query processing
   - Custom CSV agent implementation

2.2 Core Functionalities

Data Cleaning Pipeline:

```
def clean_dynamic_csv(df):
    for col in df.columns:
        if pd.api.types.is_numeric_dtype(df[col]):
            # Handle numeric data
            df[col] = df[col].fillna(df[col].mean())
            z_scores = np.abs(stats.zscore(df[col].dropna()))
            df = df[z_scores < 3]  # Remove outliers
        elif pd.api.types.is_object_dtype(df[col]):
            # Handle categorical data
            df[col] = df[col].fillna(df[col].mode()[0])
            df[col] = df[col].astype('category')
```

AI Query Processing:

```
agent = create_csv_agent(client, cleaned_file_path, verbose=True)
response = agent.run(question)
```

3. Features & Results

3.1 Key Features

1. **Automated Data Cleaning**
   - Missing value imputation
   - Outlier detection and removal
   - Data type standardization
   - Statistical validation
2. **Interactive Query Interface**
   - Natural language query support
   - Real-time response generation
   - Context-aware answers
   - User-friendly interface
3. **Data Management**
   - CSV file support
   - Cleaned data download option
   - Progress tracking
   - Error handling

3.2 Performance & Benefits

System Performance: - Efficient processing of standard CSV files - Real-time query responses - Robust error handling - Scalable architecture

Business Benefits: - Reduced data cleaning time - Improved data quality - Enhanced accessibility - Simplified analysis process

4. Future Enhancements

Short-term Improvements

1. Support for additional file formats
2. Enhanced visualization capabilities
3. Advanced query suggestions
4. Performance optimizations

Long-term Vision

1. Integration with other data sources
2. Advanced analytics features
3. Custom reporting templates
4. Machine learning model integration

5. Conclusion

The Data Query and Anomaly Removal System successfully demonstrates the potential of combining AI-powered data cleaning with natural language processing. It provides a robust foundation for data quality management while making data analysis accessible to users of all technical levels.

The system's modular architecture ensures scalability and maintainability, while its intuitive interface makes it accessible to a wide range of users. Future enhancements will further expand its capabilities and use cases across different domains.