Project 8: Excel Q&A Assistant

Objective (Why?)

Build an intelligent Excel Q&A Assistant that allows users to upload Excel files and ask analytical questions in natural language. This project combines data analysis with conversational AI, creating a powerful business intelligence tool. You will practice:

- Data Analysis with Pandas: Excel file processing and complex data manipulations
- Natural Language to Code: Converting user questions into executable Python operations

- LLM-Powered Analytics: Using AI to interpret analytical queries and generate insights
- Dynamic Visualization: Creating charts and graphs based on query results

Core Requirements (Must-have)

Component	Requirement
Excel Processing	Upload and parse .xlsx/.xls files with multi-sheet support and data validation
Natural Language Queries	Interpret business questions and convert them to Pandas operations automatically
Data Analysis Engine	Perform calculations, aggregations, and statistical analysis on Excel data
Visual Response Generation	Create charts, tables, and graphs based on query results using Plotly
Export Functionality	Download analysis results in Excel, CSV, or image formats

Milestone 1: Excel Processing & Data Intelligence

Deliverables:

- Multi-format Excel file upload and parsing system
- Automated data profiling and schema detection
- Data cleaning and preprocessing pipeline
- Interactive data preview with summary statistics
- Multi-sheet navigation and data validation

Review Requirements (Must Pass to Proceed):

- Security Review: File upload security, data validation, input sanitization

- Performance Review: Efficient Excel processing and memory management
- Code Quality Review: Clean data processing architecture

Milestone 2: Natural Language Query & Code Generation

Deliverables:

- LangChain integration for natural language processing
- Pandas code generation from natural language queries
- Query execution engine with result validation
- Interactive query interface with error handling
- Context-aware suggestions and query assistance

Review Requirements (Must Pass to Proceed):

- AI Integration Review: Sophisticated NL to code translation
- Security Review: Code execution security and sandboxing
- Performance Review: Query processing efficiency

Milestone 3: Visualization & Business Intelligence

Deliverables:

- Advanced visualization generation with Plotly
- Interactive dashboard with drill-down capabilities
- Export functionality (Excel, CSV, PNG, PDF)
- Business intelligence insights and recommendations
- Production optimization and comprehensive testing

Review Requirements (Must Pass for Project Completion):

- AI Integration Review: Intelligent data analysis and insights
- Architecture Review: Complete data analysis platform
- Performance Review: Optimized visualization and export performance
- Code Quality Review: Production-ready data analysis code

Measurable Goals & Review Template Compliance

Primary Objectives (Must Complete for Project Advancement)

- Data Analysis Mastery: Pass AI Integration Review with 9.0/10+ score (NL to code)
- Architecture Excellence: Pass Architecture Review with 8.5/10+ score
- Security Standards: Pass Security Review with 8.5/10+ score (code execution)
- Performance Optimization: Sub-5s query processing and visualization
- Code Quality Standards: Pass Code Quality Review with 8.5/10+ score

Performance Standards

- Query Processing: < 5 seconds for complex data analysis queries
- Excel Processing: < 10 seconds for typical business spreadsheets
- Visualization: < 3 seconds for chart generation and rendering
- Export Performance: < 5 seconds for multi-format exports

Task Tracking & Project Management Integration

Epic: Project 8 - Excel Q&A Assistant

Epic ID: P8-EXCEL-AI
Priority: High

Milestone 1: Excel Processing & Data Intelligence

Feature 8.1: Advanced Excel Processing System

Task ID: P8-M1-EXCEL
Priority: Critical
Sub-tasks:

- P8-M1-EXCEL-01: Multi-format Excel parser
  - Description: Comprehensive Excel file processing with multi-sheet support
  - Acceptance Criteria: Parse complex Excel files with data validation
- P8-M1-EXCEL-02: Automated data profiling
  - Description: Schema detection, data types, and statistical analysis
  - Acceptance Criteria: Complete data profile with insights
- P8-M1-EXCEL-03: Data cleaning pipeline
  - Description: Handle missing values, formatting, and data quality
  - Acceptance Criteria: Clean, analysis-ready datasets

Feature 8.2: Interactive Data Preview

Task ID: P8-M1-PREVIEW
Priority: High

Sub-tasks:

- P8-M1-PREVIEW-01: Data visualization interface
  - Description: Interactive data preview with summary statistics
  - Acceptance Criteria: User-friendly data exploration interface
- P8-M1-PREVIEW-02: Multi-sheet navigation
  - Description: Seamless navigation between Excel worksheets
  - Acceptance Criteria: Intuitive sheet selection and preview

Milestone 2: Natural Language Query & Code Generation

Feature 8.3: AI-Powered Query System

Task ID: P8-M2-AI
Priority: Critical

Sub-tasks:

- P8-M2-AI-01: LangChain integration
  - Description: Natural language processing for data queries

  - Acceptance Criteria: Accurate NL to pandas code translation
- P8-M2-AI-02: Code execution engine
  - Description: Safe pandas code execution with validation
  - Acceptance Criteria: Secure, sandboxed code execution
- P8-M2-AI-03: Query assistance system
  - Description: Context-aware suggestions and error handling
  - Acceptance Criteria: Intelligent query assistance and feedback

Milestone 3: Visualization & Business Intelligence

Feature 8.4: Advanced Analytics & Visualization

Task ID: P8-M3-VIZ
Priority: Medium

Sub-tasks:

- P8-M3-VIZ-01: Intelligent chart generation
  - Description: Automated visualization selection and creation
  - Acceptance Criteria: Context-appropriate visualizations
- P8-M3-VIZ-02: Interactive dashboard
  - Description: Business intelligence dashboard with drill-down
  - Acceptance Criteria: Professional BI dashboard interface
- P8-M3-VIZ-03: Export and sharing system
  - Description: Multi-format export with professional formatting
  - Acceptance Criteria: High-quality exports in multiple formats

Technical Specifications

Excel Data Processing

- Implement comprehensive Excel file parsing using Pandas with support for multiple worksheets
- Create automatic data schema detection and type inference for various data formats

- Build data cleaning and preprocessing pipeline to handle missing values and formatting issues

Natural Language Query Engine

- Integrate LangChain with OpenAI for converting business questions into executable Pandas code
- Design prompt engineering system for accurate query interpretation and code generation
- Implement query validation and execution framework with error handling and result verification

Visualization and Export System

- Create dynamic chart generation using Plotly with automatic chart type selection based on data types
- Build multi-format export functionality supporting Excel, CSV, PNG, and PDF outputs
- Implement interactive data tables and summary statistics display for comprehensive analysis results

Project Structure

- Build Streamlit frontend with intuitive file upload and chat-style query interface
- Create FastAPI backend for heavy data processing and analysis operations (optional for advanced users)
- Implement modular data processing pipeline with separate components for parsing, analysis, and visualization
- Design efficient session management for handling multiple files and maintaining query history

Sample User Interactions

Sales Analysis Example

```
None
User uploads: quarterly_sales.xlsx
User: "What was our total revenue in Q3?"
Assistant: "Q3 total revenue was $2,847,293. This represents a 15.3% increase from Q2."
[Shows bar chart of quarterly revenue comparison]

User: "Which product category had the highest growth rate?"
Assistant: "Electronics category had the highest growth at 23.4%, followed by Home & Garden at 18.7%."
[Displays growth rate comparison chart]
```

Financial Analysis Example

```
None
User uploads: expense_report.xlsx
User: "Show me monthly expenses breakdown as a chart"
Assistant: [Generates pie chart of expense categories by month]
"Here's your monthly expense breakdown. Marketing represents 34% of total expenses."

User: "Compare Q1 vs Q2 spending"
Assistant: [Shows comparative bar chart]
"Q2 spending increased by 12.8% compared to Q1, primarily due to increased marketing investment."
```

Advanced Features (Stretch Goals)

- Cross-Sheet Analysis: Query across multiple worksheets within the same Excel file
- Trend Detection: Automatic identification of patterns and anomalies in data

- Query Templates: Pre-built questions for common business analysis scenarios
- Data Comparison: Side-by-side analysis of different time periods or categories
- Smart Suggestions: AI-powered recommendations for relevant questions based on data structure

Deliverables

1. Complete Excel Q&A System with natural language processing
2. Data Analysis Engine supporting complex business queries
3. EXCEL_DEMO.md with sample analyses and query examples
4. GitHub Repository with comprehensive documentation
5. Live Demo showing end-to-end Excel analysis workflow

Performance Requirements

Processing Performance

- Excel file upload and parsing: < 30 seconds for files up to 10MB
- Query response time: < 10 seconds for standard analytical questions
- Chart generation: < 10 seconds for interactive visualizations

Data Handling

- Support Excel files with limited rows efficiently
- Handle multiple worksheets (up to 3 sheets per file)
- Query accuracy: nearly accurate for common business questions

Testing Scenarios

Excel Processing Testing

- Upload various Excel formats (.xlsx, .xls)
- Handle files with multiple worksheets
- Process files with different data types and formats
- Validate data cleaning and preprocessing

- Test with large files (approaching 10MB limit)

Query Intelligence Testing

- Basic numerical queries ("What is the total sales?")
- Comparative analysis ("Which month had highest revenue?")
- Growth rate calculations ("What was the YoY growth?")
- Category analysis ("Break down expenses by department")
- Trend identification ("Show me quarterly trends")

Visualization Testing

- Export functionality for charts and tables
- Large dataset visualization performance

Common Business Queries to Support

Financial Analysis

- "What was the total revenue for Q3?"
- "Show me monthly expense breakdown"
- "Which department had the highest costs?"
- "Compare this year vs last year performance"

Sales Analysis

- "What are our top 5 products by sales?"
- "Which region has the highest growth rate?"
- "Show me seasonal sales trends"
- "What is our average order value?"

Operations Analysis

- "How many orders were processed each month?"
- "What is our customer retention rate?"
- "Show me inventory turnover by category"
- "Which suppliers deliver on time most often?"

Quick Start Resources

- Pandas Excel I/O: https://pandas.pydata.org/docs/reference/api/pandas.read_excel.html

- Plotly Python: https://plotly.com/python/
- Streamlit Data Apps: https://docs.streamlit.io/knowledge-base/tutorials/build-conversational-apps
- LangChain Pandas: https://python.langchain.com/docs/integrations/tools/pandas/

Security & Data Considerations

Data Security

- Implement secure file handling with temporary storage and automatic cleanup
- Validate Excel files for malicious content and size limits
- Ensure generated Pandas code is safe and cannot execute harmful operations

Performance Optimization

- Use efficient Pandas operations and avoid memory-intensive operations
- Implement data sampling for very large datasets to maintain responsiveness
- Cache processed data and analysis results for repeated queries

FAQ

- "What Excel formats are supported?" .xlsx, .xls, and .xlsm files up to 10MB
- "How complex can the queries be?" Focus on common business analysis; complex statistical analysis is stretch goal
- "Can I analyze multiple files simultaneously?" Single file analysis is core requirement; multiple files is advanced feature
- "What if the LLM generates incorrect code?" Implement validation and fallback to predefined query patterns

Success Criteria Checklist

- Excel files upload and parse correctly with data preview
- Natural language questions convert to accurate Pandas operations
- Query results display in appropriate text and visual formats

- Charts automatically select correct visualization types
- Export functionality works for multiple formats
- Multi-sheet analysis capabilities functional
- Error handling provides helpful user feedback
- Performance meets specified response time targets
- Interface is intuitive for non-technical business users
- Query accuracy rate exceeds 90% for common questions