

# Dataset Analysis Chatbot - Work Summary

**Date:** August 29, 2025

**Project:** AI-Powered Data Analysis Chatbot

## What I Built Today

- **Dataset Upload System:** Users can drag-and-drop CSV files (up to 200MB)
- **Natural Language Chat Interface:** Ask questions about data in plain English
- **Automated Analysis:** System provides statistics, unique values, and data insights

## Key Features Working

- File upload with multiple format support
- Basic statistical analysis (min, max, mean, unique counts)
- Categorical data identification
- Geographic data filtering
- Real-time chat responses

## Current Issues

- **Accuracy Problems:** Sometimes gives wrong or incomplete answers
- **Inconsistent Results:** Response quality varies between queries

## Sample Queries That Worked

- Age range analysis (19-87 years, mean: 41.17)
- Job categories (12 unique types)
- Education levels (4 categories: primary, secondary, tertiary, unknown)
- Cricket player data analysis (Asia: 62% of 134,587 players)

## Next Steps

1. **Fix accuracy issues** - improve response validation
2. **Add advanced analytics** - charts, correlations, predictions
3. **Better error handling** - fallback for unclear queries
4. **Enhanced UI** - query suggestions and result export

The image displays two screenshots of the 'Dataset Analysis Chatbot' interface. The left screenshot shows the initial state where a dataset named 'bank.csv' (450.7KB) has been successfully uploaded. The right sidebar contains options to 'Initialize Model' and 'Settings'. The main chat area shows a user query: 'What is the age range of customers in the dataset?'. The chatbot responds with a summary of the 'age' column: Min: 19.0, Max: 87.0, Mean: 41.17, and 67 unique values.

The right screenshot shows the same interface after uploading a second dataset, 'players\_data\_with\_all\_info.csv' (4.1MB). The chatbot has processed several queries related to cricket data. For example, when asked 'what are the different types of battingstyle in dataset?', it lists 2 unique values: 'right-hand-bat', 'left-hand-bat', and 'nan'. Another query, 'which is most common bowlingstyle in the dataset?', lists 10 unique values including 'legbreak', 'right-arm-fast-medium', etc. A SQL query is also shown: `SELECT COUNT(*) FROM cricket_players WHERE country_name = 'Asia';`. The chatbot concludes by stating that the dataset contains information about 134,587 players and that approximately 62% of them are from Asia, though many have missing data for the 'asia' country\_name field.