# Problem Statement for Data Visualization Expert and AI/ML Expert Assessment

**Dataset Overview**

You have been provided with a dataset containing information about industries in India that are not licensees of BIS. The dataset includes attributes such as the name and location of industries, their geographic coordinates, contact details, scale of operations, and information related to the types of products manufactured. Some data entries may be incomplete or contain inconsistencies.

**Tasks for Data Visualization Expert**

1. **Data Cleaning & Exploratory Data Analysis (45 mins)**
   * Identify and visualize columns with missing values.
   * Present insights into which columns have the highest data incompleteness and possible reasons.
   * Suggest strategies to handle missing data.
2. **Geospatial Visualization (1 hour)**
   * Plot the geographic distribution of industries on a map of India using latitude and longitude.
   * Highlight states with the top 5 highest and lowest counts of industries.
   * Include an interactive layer that allows users to click on a state and view district-level counts.
3. **Industry Trends Analysis (1 hour 15 mins)**
   * Visualize the distribution of industries by state and scale of operation (small, medium, large).
   * Create a dashboard to show:
     + Industry count by state
     + Top 5 cities with the highest number of industries
     + Complete vs. incomplete contact information percentage
4. **Final Presentation (1 hour)**
   * Prepare and present a concise report summarizing findings with charts and maps.

Additional Tasks:

1. **Industry Growth Analysis (1 hour)**
   * Analyze and visualize industry growth trends over time using the year column.
   * Identify and visualize which states/districts show growth or decline in industrial activities.
   * Provide actionable insights based on trends.
2. **Data Completeness Dashboard (1 hour)**
   * Develop a dashboard that highlights:
     + Columns with missing data percentages
     + Counts of complete vs incomplete records for key fields (e.g., phone numbers, email IDs)
     + An interactive feature to filter by state or district and view data completeness metrics

**Important Note**: The candidates may offer further insights into dataset shared with them which may be considered for grading.

**Tasks for AI/ML Expert**

1. **Data Cleaning & Preprocessing (45 mins)**
   * Identify columns with missing data and explain strategies to fill these missing values.
   * Drop irrelevant columns (e.g., columns with near-zero variance) and justify your choices.
2. **Anomaly Detection (1 hour 15 mins)**
   * Build a model to detect anomalies in latitude and longitude values (if any) using clustering algorithms (e.g., DBSCAN).
   * Provide insights into potential data errors or fraudulent entries.
3. **Clustering & Grouping (1 hour 15 mins)**
   * Perform clustering of industries based on features like state, scale, and geolocation.
   * Visualize the clusters and explain the logic behind them.
4. **Model Evaluation & Insights (45 mins)**
   * Evaluate clustering results and present actionable insights for policymakers, such as identifying underserved industrial regions.

Additional Tasks:

1. **Predictive Model for Missing Data (1 hour)**
   * Build and train a machine learning model to predict missing values in the scale column using other relevant features like state, district, etc.
   * Evaluate the model’s performance and provide recommendations for improvement.
2. **Industrial Hotspot Prediction Model (1 hour)**
   * Develop a model to predict the likelihood of an industry being established in a particular location (state or district) based on geospatial and operational features (state, scale, products\_manufacture).
   * Explain the model choice, key features used, and insights derived from the model predictions.

**Important Note**: The candidates may offer further insights into dataset shared with them which may be considered for grading.

**Important Notes**

* Clearly document your approach, including data cleaning strategies, model choices, evaluation methods, and key insights.
* Ensure visualizations are clear, meaningful, and interactive where applicable.
* Explain the rationale behind your model selection and feature engineering techniques.
* Submit all code, visualizations, and a final report summarizing your findings.