# Healthcare Insurance Data Project:

## Mentor: Vijay Koneru

## Brief Problem Statement:

The data project aims to assist individuals in making well-informed insurance choices tailored to their needs. It does this by showcasing cost differences among insurance providers and hospitals based on location.

The healthcare industry faces considerable challenges in providing clear information about the costs associated with different insurance providers and healthcare services. Patients often struggle with a lack of transparency in pricing and data on which insurance options are most economical for specific treatments at various hospitals. The ClearCare Data Initiative seeks to address this issue by gathering, analyzing, and presenting data that compares the costs of treatments offered by different insurance providers across hospitals. Ultimately, the goal is to empower consumers with the information they need to make cost-effective healthcare decisions.

## Success Criteria:

1. **Comprehensive Data Collection**: Successfully acquiring accurate and up-to-date data from multiple hospitals and insurance providers, encompassing a wide range of treatments and services along with billing codes.
2. **Data Accuracy and Reliability**: Ensuring the data presented is accurate, reliable, and regularly updated to maintain the project's credibility and usefulness to consumers.
3. **Effective Visual Representation**: Developing clear, informative, and interactive visualizations that effectively display cost comparisons and insights for users. These visualizations should enable consumers to easily understand cost differences between insurance providers for various treatments at different hospitals.

## High-Level Requirements Summary

1. **Data Collection:**
   - Sources: Gather healthcare pricing data from hospitals and insurance providers, focusing on publicly available price transparency data, hospital charge lists, and insurance premiums.
   - Types: Include data on treatment types, hospital names, geographical locations, insurance provider names, and cost breakdowns (e.g., hospital fees, insurance coverage, out-of-pocket expenses).
2. **Data Processing:**
   - Normalization: Standardize data formats across different sources to ensure consistency in treatment names, gross charges, payer specific negotiated charges and min/max negotiated charges.
   - Cleaning and Validation: Remove duplicates, fill in missing values, and validate data integrity by checking for outliers or unrealistic cost figures.
3. **Data Analysis:**
   - Cost Comparisons: Provide comparative analysis across hospitals and insurance providers, highlighting significant cost differences for common procedures.
   - Insurance Comparison: Analyze how different insurance plans cover various treatments and the resulting out-of-pocket costs for consumers.
4. **Data Visualization:**
   - Visual Representation: Create visual representations (e.g., charts, graphs, heatmaps) to illustrate cost variations across hospitals, regions, and insurance providers.
   - User-Friendly Formats: Ensure visualizations are easy to interpret, focusing on comparative insights such as the lowest cost providers and most expensive regions.
5. **Data Documentation:**
   - Access: Ensure the dataset is available in standard formats (e.g., CSV, Excel) for easy analysis by stakeholders.

## General Guidelines on Deliverables

1. **Data Collection Framework:**

    - Data Sources List: A comprehensive list of data sources (hospitals, insurance providers, public databases) for gathering pricing information.
    - Data Collection Templates: Standardized templates for collecting data, including fields for treatment names, costs, insurance provider details, etc.

**Raw Dataset:**

- Initial Data Dump: Raw, unprocessed data collected from all sources, including treatment costs, insurance coverage, and hospital details.
- Data Dictionary: A document explaining the structure, fields, and meanings of each data point in the raw dataset.

**Cleaned and Normalized Dataset:**

- Processed Dataset: A clean and standardized dataset, ready for analysis, including normalized treatment names, consistent cost units, and resolved data gaps.

**Visual Representations** (basic story from the data collected)

- Charts and Graphs: Visual representations of the data, such as bar charts, heatmaps, and line graphs, illustrating cost variations across hospitals and insurance providers.

**Final Dataset:**

- Dataset that can be published: A finalized version of the dataset available in structured database such as PostgreSQL or user-friendly formats (e.g., CSV, Excel), cleaned, validated, and documented for ease of use.

**Project Summary and Presentation:**

- **Presentation Deck:** A slide deck summarizing the project’s objectives, process, key insights, and final deliverables for presentation to stakeholders.

## Deliverables:

1. **Project Proposal:** A document outlining the chosen website domain, the rationale for the choice, a preliminary analysis of the scope of the problem, proposed architecture, technologies to be used, and a project timeline.
2. **Interim Report:** A report detailing the progress made so far, including the data discovery, candidate healthcare provider files collected for major cities in the states assigned to your team, Preliminary observations, and any challenges encountered. Include preliminary evaluation results.
3. **Final Project Report & Presentation:** A comprehensive report covering all aspects of the project. This includes:
   - Detailed description of the files, formats, attributes, and major findings.
   - Presentation of the evaluation methodology and results.
   - Discussion of the project's limitations and future work.
   - Complete code documentation.
   - Clean/curated data in structured format: A finalized version of the dataset available in a structured database such as PostgreSQL or user-friendly formats (e.g., CSV, Excel), cleaned, validated, and documented for ease of use.

## Success Criteria:

- Effort in discovering and cleaning data: In every state assigned to your team, you will at least address two healthcare providers in every major city concerning your research.
- Completeness: will be assessed by evaluating the data team’s ability to answer questions covering different aspects of the website's content.
- Technical Quality: The raw data and cleaning procedures/transformation code should be well-structured, documented, and easily understood. The project

should demonstrate a good understanding of the chosen technologies and techniques.

- Novelty/Creativity: Projects that demonstrate innovative approaches to analyze data and understand insights from that data from your problem statement perspective would get higher recognition.

## Recommended Technologies:

- You are free to choose any open sources or readily available technical stack to data take data from raw to curated format.
- You are expected to create a structure for your final consolidated data for that state and put it in the cumulative database/data store for quick access and insights.

## Example of one hospital data record:

List of Hospitals: **https://www.hospitalsafetygrade.org/all-hospitals**

While the link above provides a comprehensive list of hospitals across the US, please select hospitals from states and institutions with which you are already familiar. This approach will streamline the process and reduce complexity.

Link to the law: https://www.cms.gov/priorities/key-initiatives/hospital-price-transparency/hospitals
The above link provides the law details for further knowledge

Link to the sample ERD: https://www.milliman.com/en/insight/price-transparency-provider-network-identification : This provide you a sample ERD so that you can understand the most important values you need from the files you download.

Sample File for Medicalcenter texas Hospital: https://medicalcentersetexas.org/price-transparency/