# CS501 - Data Science Lab - 18/8/2025

# Dialysis Facility Performance Metrics (Healthcare)

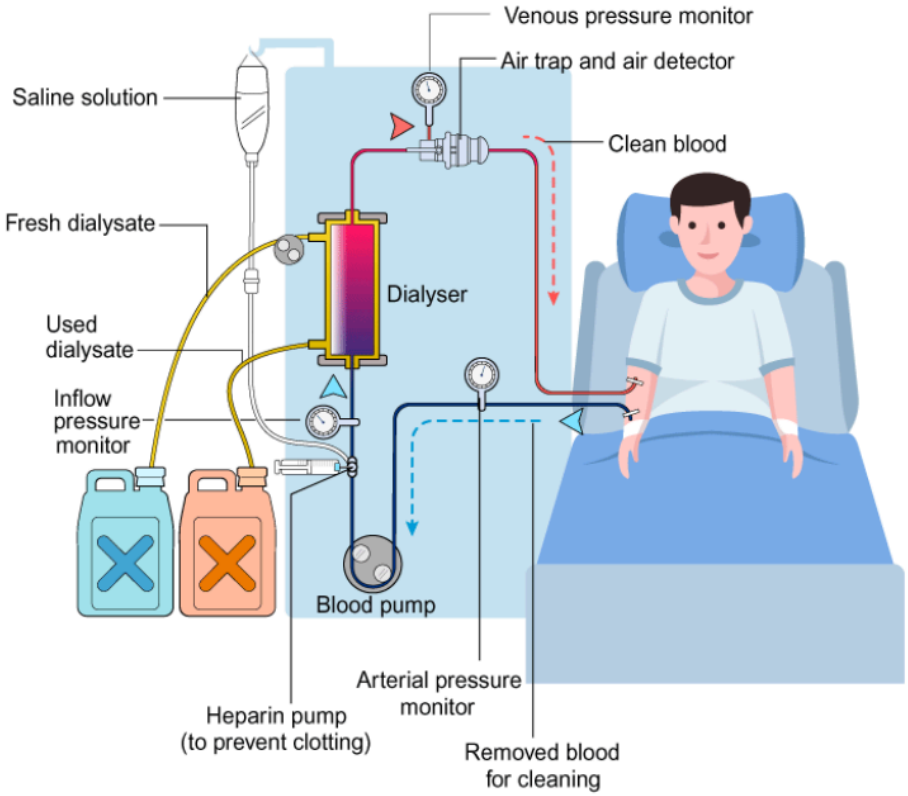
Real world Data Science is always going to have you deal with messy and unorganized data. This “conundrum” is oftentimes made even worse by unclear stakeholder expectations. Fortunately for today’s class, the data is curated very nicely and packaged.

In this Lab session you will be able to walk through it, and brainstorm about multiple ways to arrive at the solution.

The Premise -

Introductory Video on what Dialysis is and why certain people of the Society need it - https://www.youtube.com/watch?v=pAdtb1ui8YU/

You have been selected to participate in an internship for a Bioinformatics Company in the US. They have been working on a million dollar problem for a long time and want the best minds to work on it and arrive at the best solution.




The Core Problem -

In recent times, the cost of care has shot up, especially in developed countries. Kidney care, in particular dialysis, has a disproportionately high ecologic footprint within healthcare through greenhouse emissions, natural resources depletion and waste generation. Dialysis is 18 times more resource intensive than the general healthcare emissions burden of a typical patient. A typical haemodialysis session uses ~400L of tap water and ~20 kWh per patient and produces several kgs of waste. Thus, there is an urgent need to drastically lower the footprint of dialysis.

About the csv file (Dataset) -

These are lists of metrics that have been obtained from a Survey of Hospitals in the US. The Survey team told Senior Execs that it was the best data they could scrape and put together about the current status of Dialysis in the US and need your inputs to refine it.

The Variables -

A. General Information

CMS Certification Number (CCN): Unique identifier for the dialysis facility assigned by the Centers for Medicare & Medicaid Services (CMS).

Network: The network or region to which the facility belongs.

Facility Name: Name of the dialysis facility.

Five Star Date: Date range for the facility's star rating assessment.

Five Star: Star rating for the facility's overall quality (scale: 1–5).

Five Star Data Availability Code: Indicates if the five-star data is available.

B. Location Information

Address Line 1 / Address Line 2: Physical address of the facility.

City/Town: City where the facility is located.

State: U.S. state abbreviation.

ZIP Code: Postal code.

County/Parish: County or parish where the facility resides.

Telephone Number: Facility's contact number.

C. Facility Ownership and Operations

Profit or Non-Profit: Indicates the ownership type (profit or nonprofit).

Chain Owned: Whether the facility is part of a chain.

Chain Organization: Name of the parent chain, if applicable.

Late Shift: Indicates whether the facility offers late dialysis shifts.

# of Dialysis Stations: Number of available dialysis stations.

D. Treatment Types

Offers in-center hemodialysis: Whether the facility provides in-center hemodialysis.

Offers peritoneal dialysis: Availability of peritoneal dialysis.

Offers home hemodialysis training: Availability of home hemodialysis training.

E. Certification and Claims Data

Certification Date: Date when the facility was certified.

Claims Date: Date range of claims data used for assessment.

EQRS Date: Data reporting date for the End Stage Renal Disease Quality Reporting System (EQRS).

F. Mortality, Hospitalization, and Readmission

SMR Date: Date for Standardized Mortality Ratio (SMR) reporting.

Patient Survival Category Text: Survival category (e.g., "As Expected," "Better than Expected").

Patient Survival Data Availability Code: Indicates survival data availability.

Number of Patients Included in Survival Summary: Patients included in mortality analysis.

Mortality Rate (Facility): Mortality rate at the facility.

Mortality Rate: Upper/Lower Confidence Limit: Statistical confidence intervals for the mortality rate.

SHR Date: Date for Standardized Hospitalization Ratio (SHR) reporting.

Patient Hospitalization Category Text: Facility's performance in hospitalizations (e.g., "Worse than Expected").

Patient Hospitalization Data Availability Code: Indicates hospitalization data availability.

Number of Patients Included in Hospitalization Summary: Patients in hospitalization analysis.

Hospitalization Rate (Facility): Rate of hospitalizations for the facility.

SRR Date: Standardized Readmission Ratio (SRR) reporting date.

Patient Hospital Readmission Category: Performance in readmissions.

Readmission Rate (Facility): Rate of hospital readmissions within 30 days.

G. Transfusion and Transplant Metrics

STrR Date: Standardized Transfusion Ratio reporting date.
Patient Transfusion Category Text: Facility's performance in blood transfusion metrics.
Transfusion Rate (Facility): Transfusion rates at the facility.
SWR Date: Standardized Waitlist Ratio reporting date.
Standardized First Kidney Transplant Waitlist Ratio: Ratio of patients on the transplant waitlist compared to expected values.
Percentage of Prevalent Patients Waitlisted: Percentage of patients on a transplant waitlist.

H. Emergency Department and Infection Metrics

SEDR Date: Emergency Department Encounter reporting date.
Standardized ED Visits Ratio (Facility): ED visits relative to expected values.
ED30 Date: Emergency Department visits within 30 days of hospitalization reporting date.
Standard Infection Ratio (SIR): Infection rates compared to expected values.
Clinical Metrics
Fistula Rate (Facility): Percentage of patients with arteriovenous fistula for dialysis access.
HGB < 10 g/dL / HGB > 12 g/dL: Percentage of patients with specific hemoglobin levels.
Hypercalcemia: Percentage of patients with calcium levels above 10.2 mg/dL.
Serum Phosphorus Levels: Percentage of patients in specific serum phosphorus ranges.
Kt/V Data (HD/PD): Adequacy of dialysis as measured by Kt/V for hemodialysis (HD) and peritoneal dialysis (PD).

I. Staff and Patient Metrics

Healthcare Worker COVID-19 Vaccination Adherence: Percentage of healthcare workers vaccinated against COVID-19.
Long-Term Catheter Usage: Percentage of patients with long-term catheter usage.
Additional Metrics
SMoSR Date: Standardized Modality Switch Ratio reporting date.
Number of Patients in Modality Summary: Patients analyzed for switching dialysis modalities.
nPCR Data: Nutritional metrics related to dialysis adequacy for pediatric patients.

Questions for your Lab Assignment -

**EDA TASKS -**

Which states in the US are most Efficient in terms of hospital dialysis case treatment and discharge?

1. To answer this, For dialysis procedure, which state was most “efficient”?
    a. For this you would need to explore and implicate two or more variables towards “Efficiency” - which means doing the most given the least resources and producing minimal wastage.
    b. Based on the given list of Variables, Can you develop a list of variables that implicate this and justify a “sound argument” ?
    c. Answer - A para of your analysis

Hint - Read the Variables section again, and Think!
Ideally you want a birds eye view of all Variables for better understanding of the Problem.

2. Fill out the template below -
    a. My Approach to Defining "Efficiency" for Dialysis Centres -
    Define efficiency as a composite measure reflecting positive patient outcomes and lower risk of adverse events. Based on these variables - find all those variables whose higher efficiency score indicates better performance. The score will be calculated based on the aggregate of following metrics:
    b. Positive Indicators (Higher is Better):
    c. Negative Indicators (Lower is Better):
    d. Answer format - A Table of + and - indicators (minimum seven with justification)

3. Data Cleaning and Preprocessing - Then, clean the data by checking for missing values, duplicates, and any necessary transformations.
    a. Answer format - Attach the colab notebook, jupyter notebook file+screenshot.
4. Descriptive Statistics - Begin by calculating basic statistics for continuous variables (.describe() method).
    a. Answer format - Attach the colab notebook, jupyter notebook file+screenshot.

**Visualization TASKS -**

5. Visualize the Mortality Rate vs. Dialysis Stations
    a. Plot a scatter plot to see if there's any relationship between the number of dialysis stations and the mortality rate.
6. Box Plot of Mortality Rates by Facility Type

    a. Examine how mortality rates vary across different facility types (Profit vs. Non-profit).
7. Mortality Rate Confidence Interval Visualization
    a. Plot the mortality rate with its confidence interval (lower and upper bounds).
8. Comparative Analysis of Dialysis Facilities
    a. Compare the Transfusion Rate (Facility) for different regions (State).

**Think ABOUT -**

Why are you doing this Assignment? To learn more about EDA and Graphing whilst thinking about a problem critically.

Where are you likely to repeat these assignments? As a Data Scientist.

What's your goal? To mine the data, discover hidden patterns and better inform your C-suite Execs about most critical decisions they will take to drive results that can result in win-win strategy.

Marking Scheme -

EDA Part - Originality of Thinking, Clean Table of + and - correlation factors

Viz Part - Clean Graphs, Legends and Plot Accuracy.

Tips for making the visualization -

- For effective EDA visualization with Seaborn, prioritize clarity and insight.
- Choose appropriate plot types (e.g., histplot for distributions, scatterplot for relationships, boxplot for comparisons).
- Leverage hue for categorical distinctions and col for faceted views across variables. Ensure labels are clear, titles informative, and legends legible.
- Utilize Seaborn's built-in themes for aesthetic appeal and consistency.
- Focus on revealing patterns, anomalies, and relationships within the data, making visualizations interpretable and actionable.