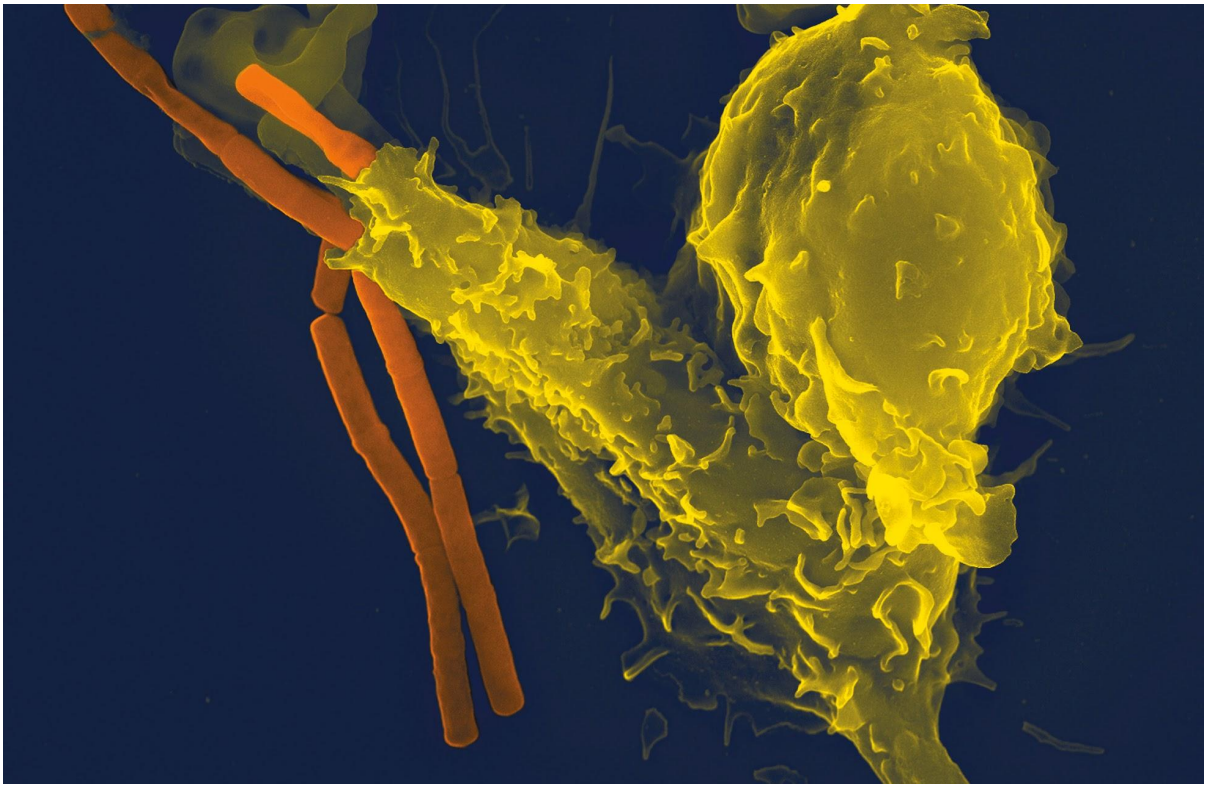

GLYCAN CLEARANCE CURVES: SITE AND CLUSTER COMPARISON

DRAFTED BY:

PRANAV SENTHILKUMARAN (ps1471)

DEPARTMENT OF STATISTICS- DATA SCIENCE

RUTGERS UNIVERSITY, NEW BRUNSWICK

ABSTRACT

.

The analysis of glycan clearance behavior is essential for enhancing the efficacy and stability of therapeutic proteins in drug development and delivery. Glycans—complex carbohydrate structures attached to proteins—significantly influence pharmacokinetics, including drug stability, efficacy, and clearance rates. This project systematically evaluates and compares glycan clearance patterns across multiple production sites (A–E) using glycan concentration data collected over specific time intervals. Key **numerical descriptors**, such as **Area Under the Curve (AUC), half-life** are calculated to quantify glycan clearance and classify them into rapid, intermediate, and slow-clearance categories.

To uncover underlying clearance patterns, **K-means clustering** is employed based on AUC and half-life values. Using the **Elbow Method**, three optimal clusters are identified:

- **Rapid-clearance glycans**: Characterized by short half-lives and low AUC, indicating limited therapeutic duration.
- **Intermediate-clearance glycans**: Display moderate clearance behavior, balancing stability and clearance.
- **Slow-clearance glycans**: Feature extended half-lives and high AUC values, contributing to prolonged therapeutic effects.

An interactive **Power BI** dashboard has been developed to provide a dynamic and visual representation of glycan behavior. The dashboard includes tools such as **stacked bar plots**, **line charts**, and **card visualizations**, enabling detailed site-specific and glycan-specific comparisons.

By integrating numerical analysis, clustering techniques, and interactive visualizations, this study delivers actionable insights into glycan clearance behavior. It establishes a robust framework for optimizing therapeutic protein design and improving drug delivery systems, thereby addressing critical challenges in pharmaceutical manufacturing and development.

KEYWORDS - Numerical descriptors, Area Under the Curve (AUC), Half-life, K-means clustering, Elbow Method, Power BI, Stacked Bar plots, Line Charts, and Card Visualizations

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Therapeutic proteins, such as monoclonal antibodies and glycoproteins, are at the forefront of modern pharmaceutical advancements. These proteins are widely used in treating various diseases, including cancers, autoimmune disorders, and infectious diseases. A critical factor influencing the effectiveness of therapeutic proteins is **glycosylation**, the enzymatic process through which glycans—complex carbohydrate chains—attach to proteins. Glycosylation is not merely a post-translational modification but a fundamental determinant of a protein's **stability, immunogenicity, clearance, and pharmacokinetics**.

Glycan clearance behavior refers to the rate and pattern by which glycans, once introduced into the biological system, are metabolized and cleared. Analyzing glycan clearance behavior is essential for optimizing therapeutic proteins because it impacts how long these proteins remain active in the body and their therapeutic efficacy. Understanding glycan behavior enables pharmaceutical researchers to design and produce more **stable, effective, and long-lasting drugs** that meet patient needs and regulatory standards.

1.1 UNDERSTANDING GLYCAN BEHAVIOUR

- **Role of Glycans in Therapeutic Proteins:**

Glycans play a multifunctional role in therapeutic protein behavior, including:

1. **Protein Stability**: Glycosylation protects proteins from enzymatic degradation, ensuring longer stability in biological systems.
2. **Pharmacokinetics**: Glycan structures influence how proteins are distributed, absorbed, and cleared from the body, directly impacting drug half-life and therapeutic duration.
3. **Immunogenicity**: Abnormal or inconsistent glycosylation can trigger immune responses, making it crucial to maintain glycan uniformity during production.

- **Key Factors Influencing Glycan Behavior:**

Glycan behavior is affected by several factors:

1. **Glycan Structure**: Differences in glycan size, branching, and composition can alter clearance rates.
2. **Production Sites**: Therapeutic proteins produced at different manufacturing sites may exhibit glycosylation variability due to changes in cell lines, culture conditions, and purification processes.
3. **Metabolic Pathways**: Glycans are cleared through specific metabolic pathways, such as hepatic or renal clearance, depending on their structural properties.

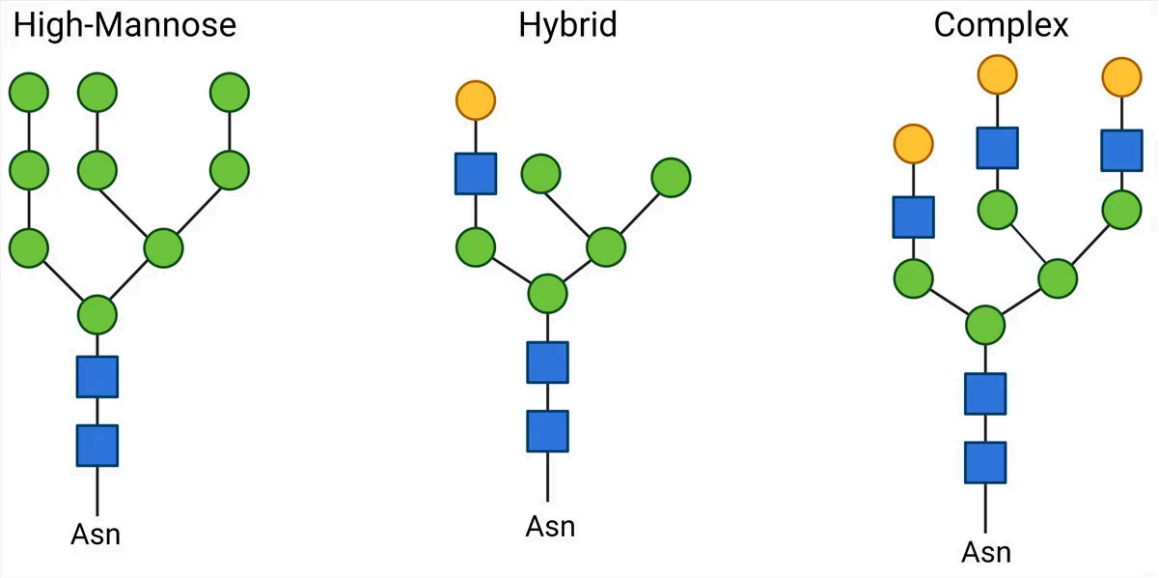

Figure 1.1 Structure of a Glycan Molecule

CHAPTER – 2

DATASET PREPARATION

Dataset preparation is a crucial step in this study, as it ensures that the glycan clearance data is clean, structured, and ready for analysis. The preparation process involves collecting raw data, cleaning and transforming it into a usable format, generating relevant numerical descriptors, and validating the quality of the processed data. This section outlines the steps undertaken for **dataset collection, preprocessing, feature engineering,** and **validation**.

2.1 DATASET PARAMETERS

The dataset includes the following core information:

1. **Site**
    - **Description:** Refers to the site of glycan measurement, with sites labeled as **A, B, C, D, and E.**
    - **Significance:** Helps differentiate the glycan clearance behaviors observed in each production site. Different sites might have varying procedures or environmental factors influencing the glycan clearance patterns.

2. **Glycan**
    - **Description:** The specific glycan molecule **(e.g., Glycan19, Glycan8).**
    - **Significance:** This parameter helps in identifying and tracking individual glycan molecules, enabling a comparison of their respective clearance rates.

3. **Count1, Count2, Count3, Count4**
    - **Description**: These columns likely represent the **number of specific monosaccharides (sugar molecules)** present in the glycan structure.
    - **Significance:** Glycans with different monosaccharide compositions might exhibit varied clearance rates, which is crucial for our analysis of their clearance behavior.

4. **Subject**
    - **Description:** Identifies the test subject or **individual** associated with the glycan measurement.
    - **Significance:** Provides insight into inter-subject variability, helping to understand how different individuals respond to glycans.

5. **Time (hours)**
    - **Description:** Time points at which glycan concentrations were measured, typically in hours.
    - **Significance:** Key to understanding the dynamics of glycan clearance over time, which allows us to calculate crucial descriptors like half-life, AUC, and clearance rate.

6. **Area Percent**
    - **Description:** Percentage of the area occupied by the glycan in chromatographic analysis.
    - **Significance:** Reflects the relative abundance of a glycan in a sample and can help in understanding how glycan concentration changes over time**.**

7. **Protein Concentration**
    - **Description:** The concentration of the protein to which the glycan is attached.
    - **Significance:** The presence of different proteins can affect the glycan behavior, thus providing context for clearance analysis.

8. **Glycan Concentration**
    - **Description:** The actual concentration of the glycan in the sample.
    - **Significance:** This is the primary variable used to calculate the glycan clearance rates, which are central to the study's objective.

2.2 DATA PREPROCESSING (WRANGLING)

We applied a rigorous data wrangling process to ensure the dataset was clean, reliable, and ready for analysis. This process included handling missing values, checking for negative or unrealistic values, removing duplicates, and transforming variables for model compatibility.

1. Checking for Missing Values:

- **Objective:** Ensure that the dataset is complete, with no missing values in critical columns like glycan concentration or protein concentration.
- **Approach:**
    - We used the **is.na()** function to identify missing values in the dataset. If any missing values were found, they were filled with appropriate placeholders (e.g., 0 for numeric values or mode for categorical values).
    - Missing values in the glycan concentration column were treated carefully, as they directly affect the analysis of clearance rates. These were imputed with the median of the respective features.

2. Handling Negative Values:

- **Objective:** Check for any negative values in columns where such values would be unrealistic or erroneous.
- **Approach:**
    - We verified that variables like Glycan Concentration and Protein Concentration did not contain negative values, as they would not make sense in this biological context. Any such instances were flagged and removed or corrected as per domain knowledge.

3. Removing Duplicate Rows:

- **Objective:** Ensure that the dataset did not contain duplicate records that could skew the analysis.
- **Approach:**
    - We applied the **duplicated()** function to identify and remove duplicate rows from the dataset. This was done across all features, with particular attention paid to entries where Subject, Glycan, and Time could potentially be repeated.

4. Final Dataset for Clustering and Model Training:

By performing these data wrangling steps, we ensured that the "***dataset was clean, transformed, and ready for the subsequent analysis and feature extraction***", which helped uncover important patterns in glycan clearance behavior.

CHAPTER 3

NUMERICAL DESCRIPTORS

In the context of glycan clearance analysis, several numerical descriptors are calculated to provide meaningful insights into the behavior of glycans over time. These descriptors summarize the essential features of glycan clearance curves and help in classifying glycans based on their pharmacokinetic properties.

The following are the primary numerical descriptors used in this study:

3.1 MEAN GLYCAN CONCENTRATION

- **Definition:**
  The mean glycan concentration represents the average concentration of a specific glycan measured over time. This statistic gives us an overall understanding of the glycan's abundance in the sample throughout the observation period.
- **Calculation:**
  It is computed as the average of glycan concentrations at all time points:

$$\text{Mean Concentration} = \frac{1}{n} \sum_{i=1}^{n} \text{Concentration}_i$$

Where:

- n is the number of time points,
- Concentration_i is the glycan concentration at the i-th time point.

Figure 3.1 Mean Concentration Calculation

- **Significance**:

The mean concentration provides a simple measure of the overall abundance of the glycan during the entire period of observation. It is useful for comparing glycans that may have different peak concentrations but exhibit similar overall behavior.

3.2 PEAK GLYCAN CONCENTRATION

- **Definition:**
  The peak glycan concentration represents the highest value of glycan concentration observed during the measurement period. This value is essential for understanding the maximum exposure of the glycan in the system.
- **Calculation:**
  The peak concentration is simply the maximum value of the glycan concentration across all time points:

$$\text{Peak Concentration} = \max(\text{Concentration}_1, \text{Concentration}_2, ..., \text{Concentration}_n)$$

Figure 3.2 Peak Concentration Calculation

- **Significance**:

Peak concentration is an important parameter for determining the maximum level of exposure to the glycan in the body. High peak concentrations might suggest a glycan that is rapidly cleared or absorbed. It provides insight into the glycan's pharmacodynamic properties and potential therapeutic implications.

3.3 AREA UNDER THE CURVE (AUC)

- **Definition:**
  The Area Under the Curve (AUC) is a crucial measure in pharmacokinetics, representing the total exposure to the glycan over time. It quantifies the cumulative glycan concentration over the measurement period, providing an estimate of the glycan's overall effect in the system.
- **Calculation:**
  AUC is computed using the **trapezoidal rule**, which approximates the area under the curve by summing up the areas of consecutive trapezoids formed between consecutive data points. The SciPy library's trapz() function is used to calculate the AUC:

Using the trapezoidal rule:

$$\text{AUC} = \int_0^T C(t)\,dt \approx \sum_{i=1}^{n-1} \frac{1}{2}\,(C(t_i) + C(t_{i+1})) \cdot (t_{i+1} - t_i)$$

Where:

- $C(t_i)$ = Glycan concentration at time t_i
- $(t_{i+1} - t_i)$ = Time interval between measurements

Figure 3.3 Area Under the Curve Calculation

- **Significance**:

AUC is a key measure of the total exposure of the glycan, as it accounts for both the concentration and the duration over which the glycan is present in the system. Higher AUC values are typically associated with prolonged therapeutic effects, which is often desired in drug design.

3.4 SLOPE

- **Definition:**
  The slope of the concentration curve measures the rate of change of glycan concentration over time. It provides insight into the speed at which the glycan concentration increases or decreases, reflecting the dynamics of glycan clearance.
- **Calculation:**
  The slope is calculated using **linear regression**.
  Specifically, **NumPy's polyfit()** function is used to fit a straight line to the concentration-time data. The slope is the first coefficient of the linear fit:

$$\text{Slope} = \frac{\Delta C}{\Delta t}$$

Where:

- ΔC = Change in concentration
- Δt = Change in time

Calculated using NumPy's polyfit():

$$\text{Slope} = \text{Coefficient of linear regression}$$

Figure 3.4 Slope of the curve Calculation

- **Significance**:

The slope helps to determine how quickly the glycan concentration changes.
A steeper negative slope indicates rapid clearance of the glycan, while a flatter slope suggests slower clearance. This is an important factor in understanding the pharmacokinetics of glycans.

3.5 HALF-LIFE

- **Definition:**
  The Half-life is the time required for the glycan concentration to reduce by half. This is a key pharmacokinetic parameter that reflects the persistence of the glycan in the system.
- **Calculation:**
  The half-life (t{1/2 }) is calculated based on the **exponential decay model**, which assumes that the concentration of a substance decays exponentially over time. The formula for half-life is:

Based on the exponential decay model:

$$t_{1/2} = \frac{\ln(2)}{k}$$

Where:

- k = Decay constant, determined from the exponential model:

$$C(t) = a \cdot e^{-kt}$$

a = Initial concentration, k = Rate constant

Figure 3.5 Half-Life Calculation

- **Significance**:

The half-life is a critical indicator of how long the glycan remains in the system. A longer half-life suggests that the glycan will stay in the system for a prolonged period, which is often desired for therapeutic applications where sustained effects are needed.

3.6 CONCLUSION

- **AUC and Half-life** are the primary numerical descriptors used for plotting results but other descriptors can also be used for fetching results according to the scenarios.
- The numerical descriptors—**mean glycan concentration, peak concentration, AUC, slope, and half-life—are key metrics** used to understand and analyze glycan clearance behaviors. These descriptors help in identifying different clearance patterns (rapid, intermediate, and slow) and provide a basis for classifying glycans according to their therapeutic potential. By combining these descriptors with clustering techniques, we can identify distinct groups of glycans that share similar clearance characteristics, which is invaluable for optimizing therapeutic protein development and drug delivery systems.

CHAPTER 4

CLUSTERING ALGORITHM IMPLEMENTATION

4.1 OBJECTIVE OF CLUSTERING

The clustering process is a vital component of the glycan clearance analysis, enabling the identification of distinct groups of glycans based on their clearance behavior. The methodology leverages numerical descriptors such as Area Under the Curve (AUC), half-life, slope, and concentration metrics to classify glycans into meaningful categories that can inform pharmaceutical manufacturing and therapeutic strategies.

The primary aim of clustering is to partition glycans into distinct groups that exhibit similar clearance patterns. By analyzing the resulting clusters, it becomes possible to:

- Identify glycans with rapid, intermediate, and slow clearance rates.
- Understand the underlying factors driving differences in clearance behavior.
- Optimize therapeutic protein development by focusing on desirable glycan clearance profiles.

4.2 FEATURE EXTRACTION

Numerical descriptors such as AUC, half-life, mean glycan concentration, and slope were computed for each glycan based on its concentration over time.

The descriptors were derived using scientific computing libraries:

- **AUC** was calculated using the trapezoidal rule with SciPy's `trapz()` function.
- **Half-life** was estimated by fitting an exponential decay model using SciPy's `curve_fit()` function.
- **Slope** was determined using linear regression with NumPy's `polyfit()` function.

These metrics summarize glycan behavior and form the foundation for clustering.

4.3 NORMALIZATION

To ensure comparability across features with different scales, all numerical descriptors were normalized using **z-score normalization: $Z=(X-\mu)/\sigma$;** where X is the feature value, μ is the mean, and σ is the standard deviation.

Normalization prevents features with larger magnitudes (e.g., AUC) from disproportionately influencing the clustering process.

4.4 ALGORITHM SELECTION

K-Means Clustering:

- **K-means clustering** is an unsupervised machine learning algorithm used to partition a dataset into KKK distinct, non-overlapping clusters. It groups data points based on their similarities by minimizing the variance within clusters while maximizing the variance between clusters.
- **K-Means** was chosen for its simplicity, computational efficiency, and ability to handle high-dimensional numerical data.
- The algorithm partitions data points into K clusters by minimizing within-cluster variance (inertia) and maximizing inter-cluster separation.

Optimal Number of Clusters:

- **ELBOW METHOD:** The Elbow Method is a popular technique used to determine the optimal number of clusters (K) in clustering algorithms such as K-means. It helps to select the number of clusters that best fit the data by evaluating the within-cluster sum of squares (WCSS), also known as the inertia.
- The **Elbow Method** was used to determine the optimal value of K. This method involves plotting the sum of squared distances (inertia) for different values of K and identifying the "elbow point" where the rate of inertia reduction decreases significantly.
- Based on the Elbow Method, **K=3** was selected, resulting in three clusters representing rapid, intermediate, and slow-clearance glycans.

4.5 CLUSTER EXECUTION

The clustering process involved the following steps:

1. **Initialization**
    - The algorithm randomly initialized KKK cluster centroids in the feature space.
2. **Assignment Step**
    - Each glycan was assigned to the nearest cluster centroid based on the Euclidean distance:

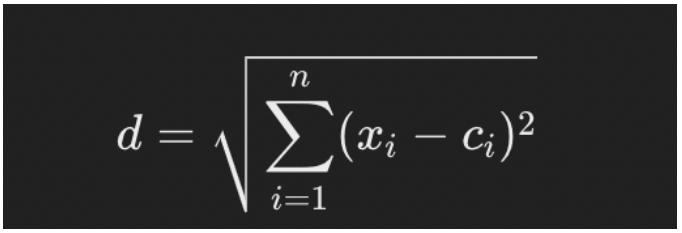

Figure 4.3 Euclidean Distance Calculation

where xi is the feature vector of the glycan and ci is the centroid of the cluster.

3. **Update Step**
    - The centroids were recalculated as the mean of all data points assigned to each cluster.

4. **Iteration**
    - The assignment and update steps were repeated until convergence, i.e., when cluster memberships no longer changed or the reduction in inertia fell below a predefined threshold.

4.6 CLUSTER ANALYSIS

Cluster analysis revealed meaningful insights into glycan clearance patterns by grouping glycans into clusters based on their pharmacokinetic properties. The analysis focused on understanding how glycans behave over time and their

potential therapeutic implications. Each cluster was characterized by its unique features derived from the numerical descriptors **(e.g, AUC, half-life, slope)**.

Cluster Characteristics

1. **Cluster 0 (Rapid Clearance):**
    - **Descriptors:**
        - Short half-life, indicating quick elimination from the system.
        - Low Area Under the Curve (AUC), suggesting minimal exposure over time.
        - Steep negative slope, reflecting a rapid decline in glycan concentration.
    - **Implications:**
        - These glycans are quickly cleared and may have limited therapeutic efficacy due to their short duration of action.
2. **Cluster 1 (Intermediate Clearance):**
    - **Descriptors:**
        - Moderate half-life, AUC, and slope values.
        - Balanced clearance rate, indicating a mix of stability and clearance.
    - **Implications:**
        - Represents a middle ground, where glycans have reasonable therapeutic exposure and clearance dynamics.
3. **Cluster 2 (Slow Clearance):**
    - **Descriptors:**
        - Long half-life, signifying prolonged presence in the system.
        - High AUC, reflecting substantial exposure over time.
        - Flat slope, indicating slow and steady clearance.
    - **Implications:**
        - Ideal for therapeutic purposes, as these glycans provide sustained effects and prolonged efficacy.

CHAPTER 5

RESULTS AND DISCUSSION

A CSV file was generated showcasing the results of glycans.

The clustering and analysis of glycan clearance data yielded meaningful insights into glycan pharmacokinetics, offering significant implications for therapeutic protein development and manufacturing processes. By categorizing glycans into clusters based on their clearance behavior, the study highlights the variability in glycan dynamics across production sites and emphasizes the importance of optimizing glycan profiles for desired therapeutic effects.

5.1 CLUSTER INSIGHTS

1. **Cluster 0 (Rapid Clearance):**
    - **Characteristics:**
        - Glycans in this cluster exhibited steep declines in concentration over time.
        - They had the shortest half-lives, indicating quick elimination from the system.
        - The lowest AUC values suggest minimal exposure during the observation period.
    - **Implications:**
        - These glycans may lead to shorter therapeutic durations, making them less favorable for long-term efficacy.
        - Rapid-clearance glycans could indicate instability or undesirable interactions within the biological system.
        - These glycans are potential candidates for process refinement to minimize their presence in production batches.
2. **Cluster 1 (Intermediate Clearance):**

- **Characteristics:**
  - Glycans in this cluster showed moderate declines in concentration.
  - Balanced half-life and AUC values indicate a mix of stability and clearance efficiency.
- **Implications:**
  - Intermediate-clearance glycans strike a balance between therapeutic exposure and clearance.
  - They may be suitable for therapies requiring moderate-duration effects.
  - This cluster provides insights into optimizing glycan profiles for specific therapeutic goals that require both efficacy and controlled clearance.

3. **Cluster 2 (Slow Clearance):**
   - **Characteristics:**
     - Glycans in this group demonstrated gradual declines in concentration over time.
     - The highest AUC values indicated prolonged exposure, and the longest half-lives suggested sustained therapeutic effects.
   - **Implications:**
     - Slow-clearance glycans are highly desirable for therapeutic applications, as they ensure prolonged efficacy and reduced dosing frequency.
     - These glycans are prime targets for optimizing production processes to increase their proportion in batches.

5.2 SITE COMPARISONS

Variation Across Sites:

- Sites A and E exhibited a higher proportion of slow-clearance glycans, which may reflect differences in manufacturing processes, environmental conditions, or starting materials.

- Other sites (e.g., B, C, and D) had higher proportions of rapid-clearance glycans, suggesting areas for potential improvement in production protocols.

Implications of Variability:

- The observed site-specific differences underscore the importance of maintaining consistency in manufacturing processes.
- Variability in glycan profiles across sites could impact therapeutic outcomes and product uniformity.
- Addressing site-based discrepancies can lead to more predictable and reliable therapeutic effects.

CSV FILE:

glycan_clearance_curves

Site	Glycan	Time (hours)	Concentration	Cluster	Mean Concentration	Peak Concentration	AUC	Slope	Half-life
A	Glycan19	1	2712.5	0	1654.8527777777800	3213.0	-327773.40000000000	-10.599335957093800	61.91654471612780
A	Glycan19	3	1849.2	0	1654.8527777777800	3213.0	-327773.40000000000	-10.599335957093800	61.91654471612780
A	Glycan19	5	2244.0	0	1654.8527777777800	3213.0	-327773.40000000000	-10.599335957093800	61.91654471612780
A	Glycan19	24	1232.0	0	1654.8527777777800	3213.0	-327773.40000000000	-10.599335957093800	61.91654471612780
A	Glycan19	72	796.4	0	1654.8527777777800	3213.0	-327773.40000000000	-10.599335957093800	61.91654471612780
A	Glycan19	168	440.8	0	1654.8527777777800	3213.0	-327773.40000000000	-10.599335957093800	61.91654471612780
A	Glycan19	1	2609.6	0	1654.8527777777800	3213.0	-327773.40000000000	-10.599335957093800	61.91654471612780
A	Glycan19	3	1876.8	0	1654.8527777777800	3213.0	-327773.40000000000	-10.599335957093800	61.91654471612780

Figure 5.2 Output CSV file representation for Site A

CHAPTER 6

DATA VISUALIZATION AND DASHBOARD WALKTHROUGH

6.1 UTILIZING POWER BI

Power BI is a powerful business analytics tool developed by Microsoft that enables users to visualize data and share insights interactively. With its user-friendly interface and robust integration features, Power BI facilitates the creation of comprehensive dashboards by seamlessly connecting to various data sources. These dashboards allow stakeholders to explore data dynamically, uncover trends, and make informed decisions based on visual analytics.

For this project, Power BI is utilized to **analyze glycan clearance data and present key metrics using line charts, stacked bar charts, and card visualizations.** Each visualization was tailored to highlight specific aspects of glycan behavior, making it easier to interpret the results and draw actionable insights.

6.2 DASHBOARD CREATION

1. Importing the Dataset:

- **Imported the output glycan clearance csv file into Power BI using its "Get Data"** feature. Power BI supports multiple file formats, including CSV, Excel, and direct database connections.
- The dataset was cleaned and processed in Python before being uploaded to ensure accuracy and consistency.

2.Data Modeling and Preparation:

- Power BI's data modeling capabilities allowed me to structure and format the data effectively.

- Relationships between columns (such as glycan, site, and time) were established using Power BI's relationship view, ensuring seamless filtering and aggregation across visualizations.

3.Visualizations Incorporated:

3.1) Line Chart: Glycan Clearance Curves

Purpose:

- Line charts were used to visualize glycan concentrations over time for individual glycans.
- These plots highlight the dynamic behavior of glycans, such as rapid or slow clearance.

How It Was Built:

- The **X-axis** represents time (hours).
- The **Y-axis** depicts the Sum of glycan concentration.
- Each glycan was assigned a unique color to differentiate its clearance curve.
- Tooltips were enabled to display exact glycan concentration values when hovering over data points.

Insights Gained:

- Rapid-clearance glycans exhibited steep declines, while slow-clearance glycans showed more gradual reductions.
- This chart provided a direct visual representation of clearance rates for each glycan, aiding in cluster categorization.

3.2) Stacked Bar Chart: AUC Distribution by Cluster

Purpose:

- Stacked bar charts were used to compare glycan behavior across different clusters and production sites (A–E).

- The Area Under the Curve (AUC) values were categorized into three clusters: rapid, intermediate, and slow clearance.

How It Was Built:

- The **X-axis** represents glycan clusters (Cluster 0, 1, and 2).
- The **Y-axis** depicts the AUC values, aggregated by production site.
- A stacked format allowed for clear visualization of site-wise contributions to each cluster.

Insights Gained:

- Cluster 0 contained glycans with low AUC, indicating rapid clearance.
- Cluster 2 showed the highest AUC values, reflecting slow clearance.
- Sites A and E exhibited a higher proportion of slow-clearance glycans, suggesting potential differences in manufacturing processes.

3.3) Card Visualizations: Key Metrics

Purpose:

- Card visualizations were used to summarize critical numerical metrics for a quick overview of glycan behaviors.
- These visuals highlighted high-level statistics such as the total number of glycans, mean AUC, and average half-life.

How It Was Built:

- Individual metrics were calculated using Power BI's **measure functions** and displayed as cards.
- Each card included a descriptive title (e.g, "Count of Glycans" or "Average Half-Life") and a dynamically updated value.

Insights Gained:

- The cards provided instant access to summary statistics, reducing the need to explore raw data for high-level information.
- Metrics like the average half-life and total glycans supported stakeholders in understanding the overall behavior of the dataset.

3.4) Interactivity Features

1. **Slicers and Filters**:
   - Added slicers to allow users to filter data by production site, glycan type, or time intervals.
   - This interactivity made it easy to focus on specific subsets of data without modifying the dashboard structure.
2. **Dynamic Legends**:
   - Each chart included legends for intuitive identification of clusters, glycans, or sites.
   - The stacked bar chart legend showed site-specific contributions to each cluster.
3. **Drill-Down Functionality**:
   - Users could drill down into specific clusters or sites to explore detailed glycan behaviors.

6.3 VISUALIZATIONS USED IN THE DASHBOARD

PLOT1: Glycan Clearance Curves (Concentration vs. Time)

- **Purpose**: This plot visualizes the glycan concentration over time, showcasing the clearance behavior of individual glycans.
- **Features**:
  - A line chart for each glycan's concentration over the observation period.
  - Distinct patterns highlight the varying clearance rates among glycans.

- **Key Findings**:
    - Glycan 11 and Glycan 10 exhibited steep declines, indicating rapid clearance.
    - Glycan 18 and Glycan 14 showed gradual reductions, reflecting slow clearance.
    - Variance in clearance behavior is apparent, with some glycans showing sharp reductions and others maintaining a prolonged presence in the bloodstream.
- **Implications**:
    - Rapid-clearance glycans, such as 11 and 10, may result in shorter therapeutic effects, which could necessitate modifications in manufacturing to enhance stability.
    - Slow-clearance glycans, like 18 and 14, exhibit potential for long-lasting therapeutic applications, making them ideal candidates for extended drug delivery.

PLOT 2: Area Under the Curve (AUC) Distribution by Cluster

- **Purpose**: This plot provides a clustered analysis of glycans based on their AUC values, enabling differentiation between rapid, intermediate, and slow-clearance behaviors.
- **Features**:
    - A bar chart or clustered histogram, grouping glycans into three clusters derived from K-means clustering.
    - Cluster 0 (**low AUC**): Represents glycans that clear quickly and have shorter exposure times.
    - Cluster 1 (**moderate AUC**): Indicates glycans with intermediate clearance behavior.
    - Cluster 2 (**high AUC**): Reflects glycans with prolonged exposure, demonstrating slow clearance.
- **Key Findings**:
    - Cluster 0 glycans are ideal for short-term effects but may not sustain therapeutic efficacy.

- Cluster 2 glycans show the longest exposure times, making them more suitable for therapies requiring extended duration.
- Cluster 1 glycans balance clearance and stability, potentially serving versatile therapeutic needs.

6.4 DASHBOARD VISUALIZATIONS OF VARIOUS SITES AND CLUSTERS

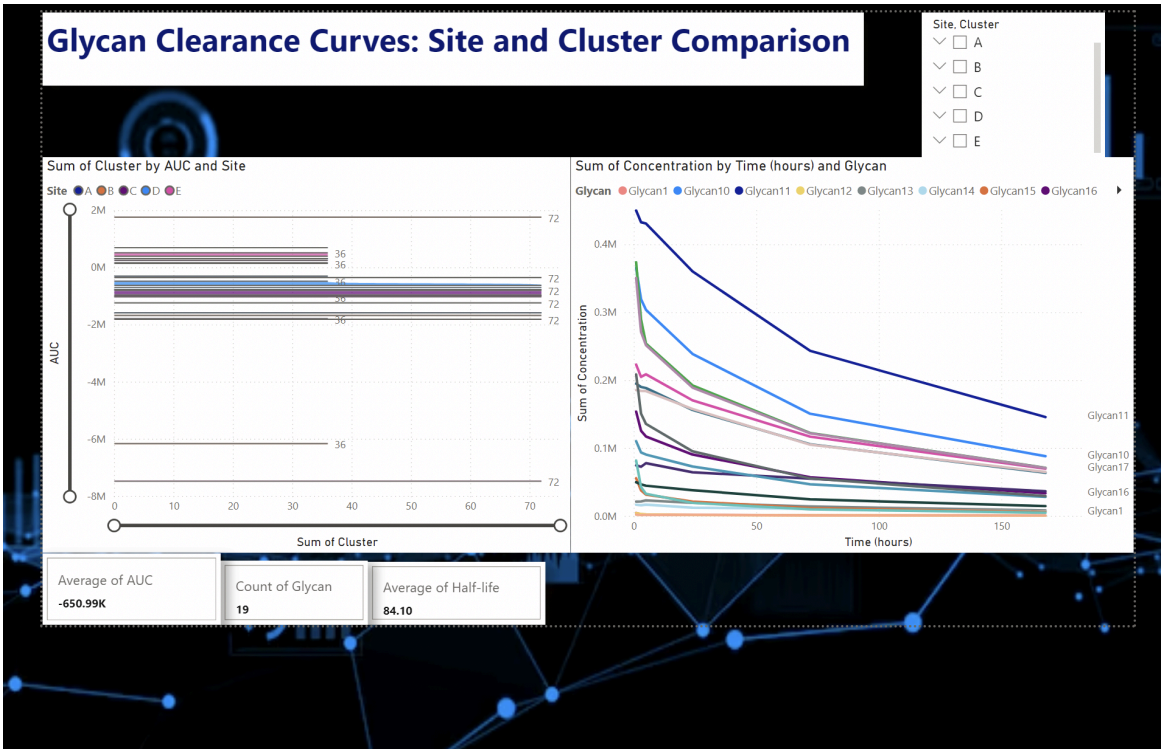

Figure 6.3.1 Overall Result (Stacked Bar chart and Line chart)

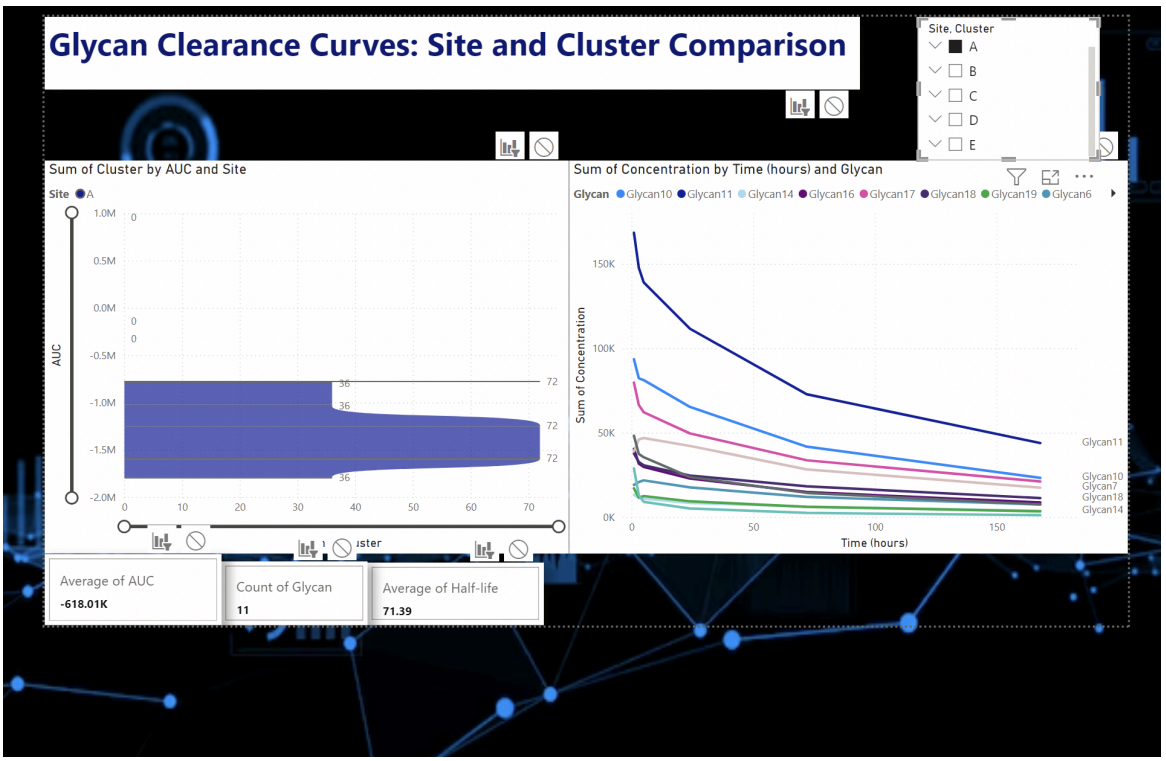

Figure 6.3.2 Cluster 0,1,2 Results of Site A

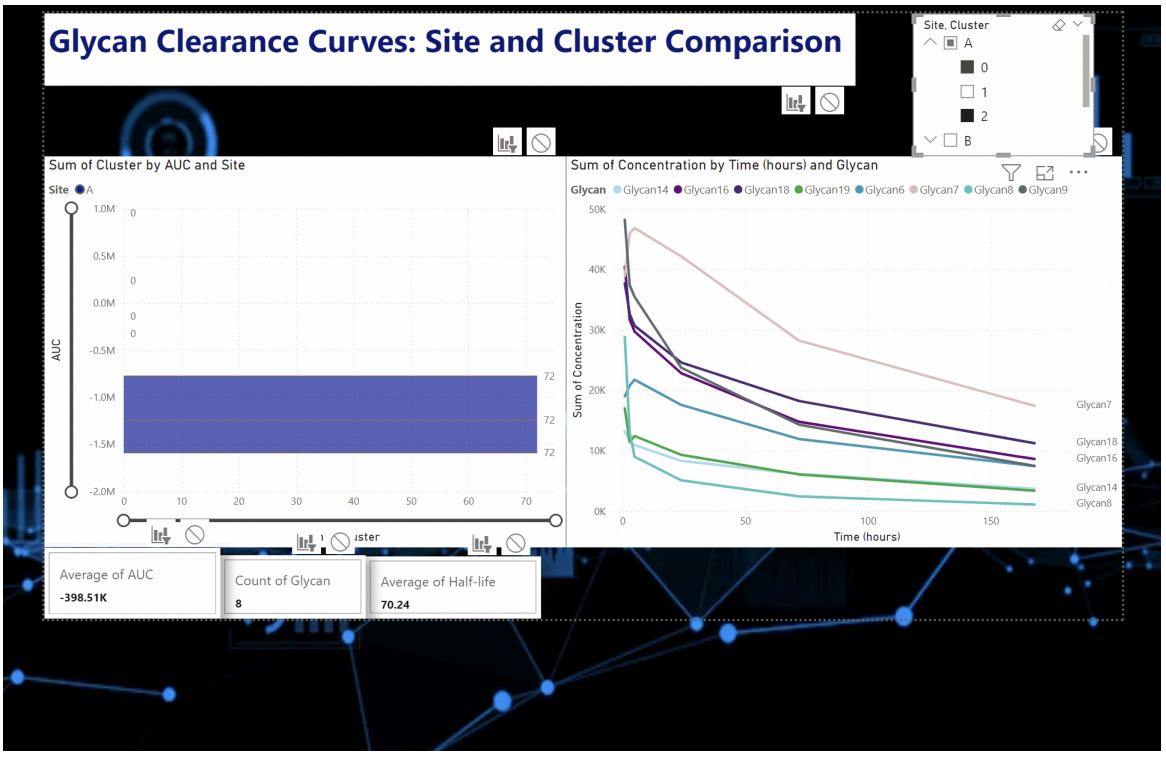

Figure 6.3.3 Cluster 0,2 Results of Site A

- **Likewise, we can check for all the possibilities which allow stakeholders to identify patterns, make data-driven decisions, and optimize glycan production for therapeutic applications.**

6.5 CONCLUSION

Using Power BI, ***an intuitive and interactive dashboard that transforms complex glycan clearance data into actionable insights is created***. The combination of line charts, stacked bar charts, and card visualizations provided a comprehensive view of glycan behavior across different sites and clusters. These visualizations allowed stakeholders to identify patterns, make data-driven decisions, and optimize glycan production for therapeutic applications. The dashboard serves as a powerful tool for bridging the gap between data analysis and real-world application in pharmaceutical development.

CHAPTER 7

APPLICATIONS AND FUTURE SCOPE

7.1 APPLICATIONS

1. **Drug Development:**
   Optimize therapeutic biologics by tailoring glycan structures to improve efficacy, stability, and patient outcomes.
2. **Manufacturing Optimization:**
   Standardize production to minimize batch variability and ensure consistent product quality.
3. **Personalized Medicine:**
   Adjust dosing based on individual glycan clearance rates to enhance treatment precision and reduce side effects.

4. **Regulatory Compliance:**
   Support quality control and meet safety standards by identifying and quantifying clearance patterns.
5. **Glycoengineering:**
   Design glycan modifications to achieve targeted pharmacokinetics for advanced therapies.
6. **Biomarker Discovery:**
   Use glycan clearance patterns as indicators for disease monitoring and therapy effectiveness.

7.2 FUTURE SCOPE OF GLYCAN RESEARCH

- **Advanced Predictive Models:**

  As more data is collected and analyzed, machine learning algorithms such as K-means clustering can be enhanced by incorporating more advanced predictive modeling techniques, including deep learning and reinforcement learning. These models could predict not only glycan clearance but also its interaction with other biologics, tissues, or the immune system. By improving the predictive power of these models, it may become possible to forecast the long-term outcomes of glycan therapies, including adverse reactions or efficacy variations across patient groups.

- **Cross-disciplinary Integration:**

  Combining glycan clearance studies with genomics, proteomics, and metabolomics data will allow for a more holistic understanding of the underlying biological mechanisms that influence glycan behavior. The integration of these fields could uncover new insights into how genetic variations or metabolic processes influence glycan clearance, opening up new avenues for personalized treatments.

- **Real-time Monitoring and Dynamic Adjustments:**

  In the future, real-time monitoring of glycan clearance could become a standard part of biologic therapy. Through advanced sensors or biomarkers, drug clearance could be tracked continuously in patients, enabling dynamic adjustments to dosage or treatment plans. This could significantly enhance the precision and personalization of biologic therapies, especially for chronic conditions that require long-term management.

CHAPTER 8

REFERENCES

8.1 RESEARCH PAPERS

- **Wang, L. X., Tong, X., & Li, C. (2019).**
  Glycoengineering of therapeutic glycoproteins: From structural biology to practical applications.
  Nature Reviews Drug Discovery, 18(2), 115-130.
  https://doi.org/10.1038/s41573-018-0005-0
- **Narimatsu, Y., Joshi, H. J., Nason, R., et al. (2019).**
  An atlas of human glycosylation pathways enables display of the human glycome by gene engineered cells.
  Molecular Cell, 75(2), 394-407.e5.
  https://doi.org/10.1016/j.molcel.2019.05.017

8.2 WEBSITES FOR FURTHER EXPLORATION

- **GlyGen Database:**

  A glycan-focused bioinformatics resource.

  https://www.glygen.org

- **PowerBI for Data Visualization:**

  https://powerbi.microsoft.com/

CHAPTER 9

SUMMARY OF WORK FLOW

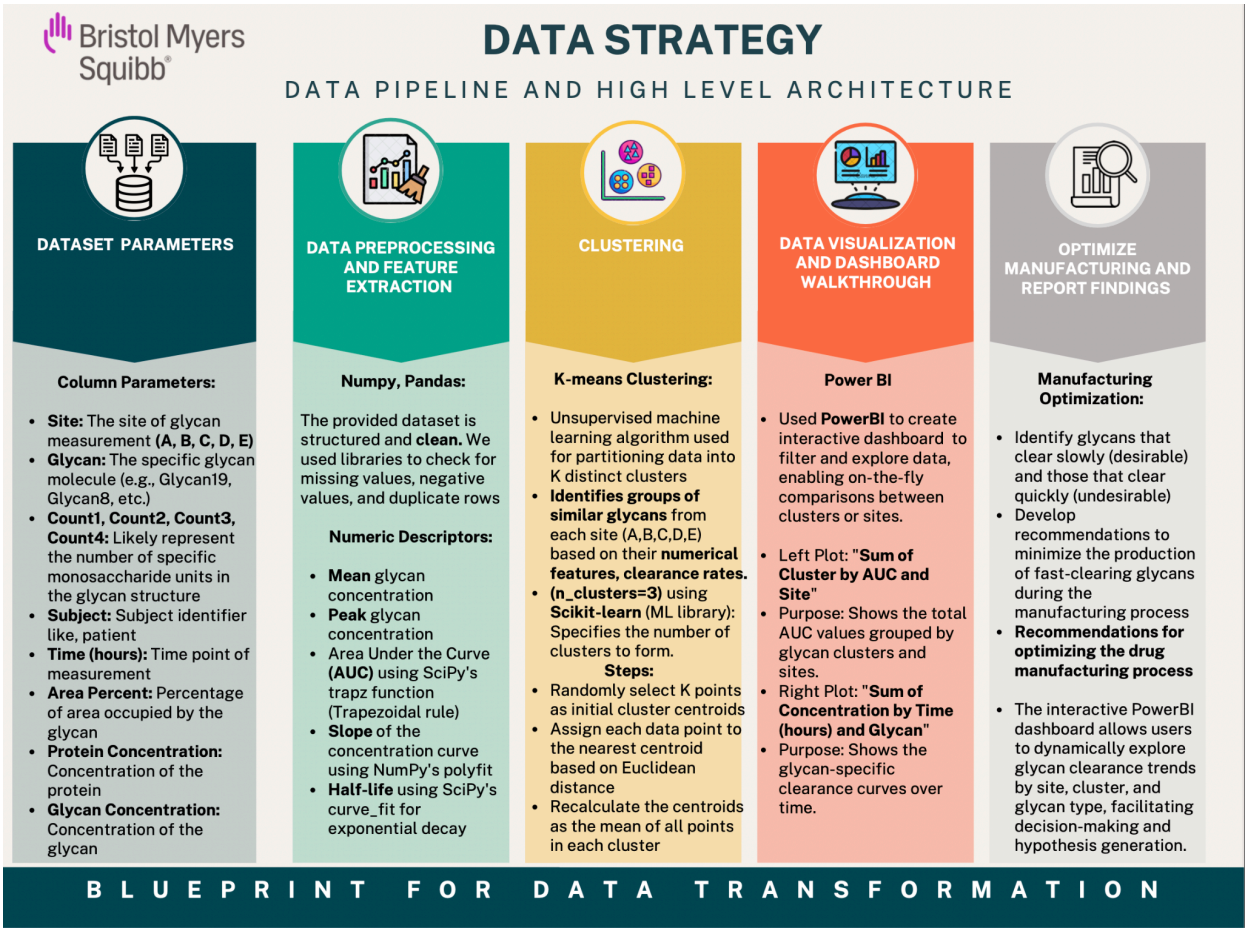

Figure 8.1 Methodology Flowchart