Question for Week 2 - submission

Your Module 1 discussion post identified some high-level goals for working with a dataset of interest to you. In this post, you will expand on those goals to characterize your target problem and develop some low-fidelity prototypes for working with that data. First, identify two to three tasks you would wish to complete with your data, identifying:

* Why is a task pursued? (goal)
* How is a task conducted? (means)
* What does a task seek to learn about the data? (characteristics)
* Where does the task operate? (target data)
* When is the task performed? (workflow)
* Who is executing the task? (roles)

Then, sketch a set of preliminary low-fidelity prototypes for addressing these tasks with the given data. You may either sketch freeform or use the Five Design Sheets approach to generate these prototypes (hand-sketched on paper is fine). Upload a copy of your sketches as part of your post.

**Characterizing the Target Problem with the Kaggle Heart Failure Clinical Records Dataset:**

The **Heart Failure Clinical Records Dataset** contains patient records related to heart failure, with a key target variable being **DEATH\_EVENT** (whether the patient survived or died). The overall goal is to analyze this dataset to understand the factors contributing to heart failure outcomes, and ultimately, to predict patient mortality.

**Goals:**

The primary goal is to identify and characterize the key factors that influence patient survival or death in heart failure cases. We aim to derive actionable insights from the data to help predict future patient outcomes and potentially inform clinical decision-making.

**Task 1: Exploring the Relationship Between Ejection Fraction and Mortality**

* **Why (Goal)**: Ejection fraction is a well-known indicator of heart function. The goal of this task is to determine whether a low ejection fraction is strongly associated with patient mortality, which could be crucial for identifying patients at high risk.
* **How (Means)**: This task is conducted through **exploratory data analysis** techniques like **box plots**, **scatter plots**. These visualizations help in examining the distribution of ejection fraction values and their relationship with mortality.
* **What (Characteristics)**: This task seeks to learn about the distribution of ejection fraction in the patient population and how it differs between patients who survived and those who did not (DEATH\_EVENT). It also aims to identify threshold values below which the risk of death increases significantly.
* **Where (Target Data)**: The task focuses on the **ejection fraction** variable and its relationship with the binary outcome **DEATH\_EVENT**. Additional variables like age, sex, and serum creatinine may also be included to understand if ejection fraction interacts with other factors.
* **When (Workflow)**: This task is usually part of the **early exploratory analysis**, where visualizations and statistical summaries are used to generate hypotheses about the data. Once relationships are identified, more complex modeling tasks may follow.
* **Who (Roles)**:
  + **Data Analyst**: Responsible for generating visualizations and summary statistics to uncover relationships.
  + **Medical Researcher**: Provides insights into the clinical relevance of observed relationships, helping to interpret the findings.

**Task 2: Identifying Demographic Risk Factors (Age, Gender, and Smoking) for Heart Failure Mortality**

* **Why (Goal)**: This task is pursued to explore how demographic factors like age, gender, and smoking status affect heart failure outcomes. Understanding demographic risk factors helps in targeted prevention and treatment strategies.
* **How (Means)**: The task is conducted through **statistical analysis** and **visualizations** like **histograms**. These tools help reveal demographic trends and compare survival rates across different subgroups.
* **What (Characteristics)**: This task seeks to learn whether certain demographic groups (e.g., older patients, smokers, males) are more prone to heart failure mortality. It also explores how these factors interact with clinical variables like blood pressure or serum creatinine.
* **Where (Target Data)**: The task operates on the **age**, **sex**, and **smoking** columns, and examines their relationship with the target outcome, **DEATH\_EVENT**. Other clinical variables may be included to control for potential confounders.
* **When (Workflow)**: This task is performed as part of both the **exploratory analysis** and during the **feature engineering** phase, where demographic variables are analyzed for their importance and potential inclusion in predictive models.
* **Who (Roles)**:
  + **Data Scientist**: Responsible for statistical analysis and visualization of demographic factors.
  + **Public Health Expert**: Can help in understanding population-wide implications of the results and suggest areas for targeted interventions.

**Conclusion:**

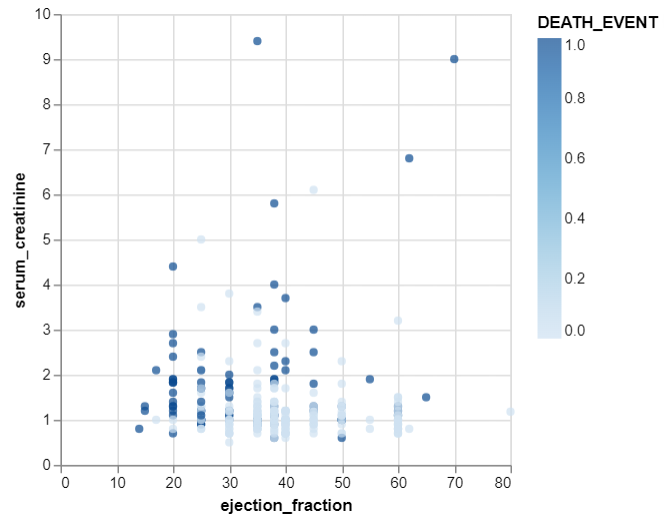
These tasks form the foundation of analyzing the **Heart Failure Clinical Records** dataset. They are conducted with the goal of understanding key predictors of heart failure mortality, focusing on critical clinical factors like ejection fraction, demographic factors, and patient-specific variables. This knowledge can be used to create predictive models that guide treatment and improve patient outcomes. Each task leverages different analytical tools, roles, and workflows, depending on the specific characteristics and insights being sought from the data.

Preliminary Low-fidelity prototypes

These prototypes represent early sketches or conceptual designs for the types of charts and dashboards that would help answer the identified tasks.

**Prototype 1: Mortality Prediction (Scatter Plot with Decision Boundary)**

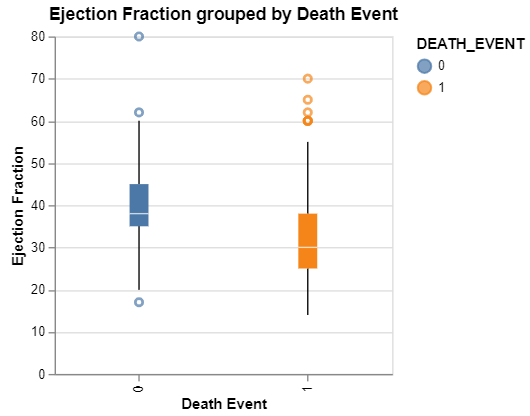
* **Purpose**: Visualize patient data points with predicted outcomes (death or survival) and the decision boundary for the prediction model.
* **Description**:
  + A **scatter plot** showing patients categorized by their clinical features like **ejection fraction** and **serum creatinine** with color indicating **predicted outcome** (e.g., red for death, blue for survival).
  + A decision boundary to visually separate high-risk and low-risk patients based on clinical data.
* **Prototype Sketch**:



**Task 2: Exploring the Relationship Between Ejection Fraction and Mortality**

**Prototype 3: Box Plot of Ejection Fraction by Mortality**

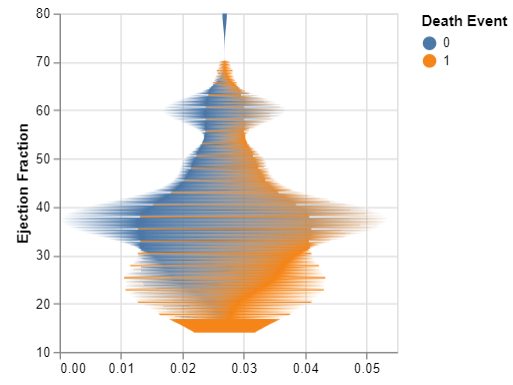
* **Purpose**: Compare the distribution of **ejection fraction** between patients who survived and those who died.
* **Description**:
  + A **box plot** showing the distribution of ejection fraction values for both groups (**survived** vs **died**). This helps in identifying whether low ejection fraction values are correlated with a higher mortality rate.
* **Prototype Sketch**:



+---------------------------------------------------+

**Prototype 4: Violin Plot of Ejection Fraction by Mortality**

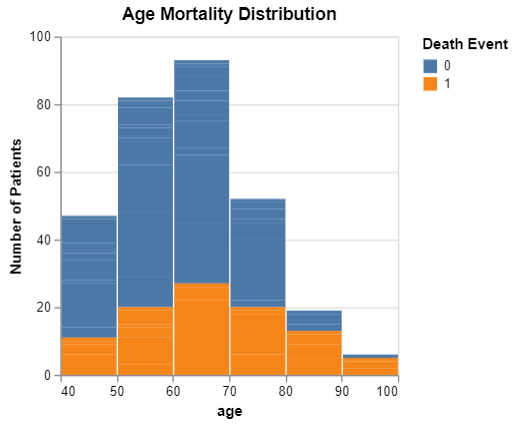
* **Purpose**: Show a more detailed distribution of **ejection fraction** with kernel density estimation, allowing users to see the distribution's shape.
* **Description**:
  + A **violin plot** comparing the distribution of ejection fraction values for patients who died vs those who survived, showing not just quartiles but the overall density of values.
* **Prototype Sketch**:

+---------------------------------------------------+

**Task 3: Identifying Demographic Risk Factors (Age, Gender, and Smoking) for Heart Failure Mortality**

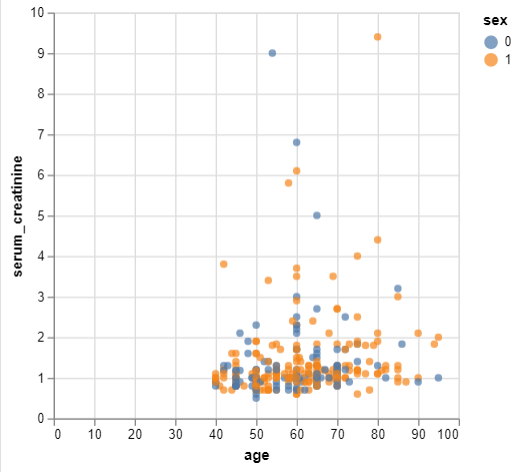
**Prototype 5: Age and Mortality Relationship (Histogram/Bar Chart)**

* **Purpose**: Examine the distribution of **age** between patients who survived and those who died.
* **Description**:
  + A **histogram or bar chart** to show the frequency of patients in different age groups (e.g., 40-50, 51-60, etc.) and how these groups correlate with mortality. The goal is to spot trends like older age being associated with higher mortality.
* **Prototype Sketch**:

+---------------------------------------------------+

**Prototype 6: Scatter Plot with Transparency (Age vs Serum Creatinine and Gender)**

* **Purpose**: Visualize the relationship between **age** and **serum creatinine** levels by **gender**, using transparency to handle data overlap.
* **Description**:
  + A **scatter plot** where **age** is on the x-axis, and **serum creatinine** is on the y-axis. The points are colored by **gender** and transparency is applied to reveal dense clusters of data points.
* **Prototype Sketch**:

+---------------------------------------------------+

**Summary of Low-Fidelity Prototypes:**

These initial low-fidelity prototypes are conceptual designs that illustrate how the given tasks might be addressed. Here's a brief recap:

1. **Mortality Prediction Scatter Plot**: Visualizes high-risk vs. low-risk patients using clinical features like ejection fraction and serum creatinine.
2. **Box Plot of Ejection Fraction by Mortality**: Compares ejection fraction distribution between patients who survived and those who died.
3. **Violin Plot of Ejection Fraction by Mortality**: Provides a detailed distribution comparison with a density view.
4. **Age Distribution by Mortality Bar Chart**: Shows the correlation between age groups and mortality.
5. **Scatter Plot with Transparency (Age vs Serum Creatinine)**: Visualizes relationships between age and serum creatinine levels with gender differentiation.

These prototypes serve as a foundation for more detailed, interactive visualizations and dashboards, providing insights into patient mortality and risk factors in heart failure.