Business Analytics

Data Collection and Preparation for Anderson Cancer Center

Presented by: Pharaoh Kipkirui Chirchir

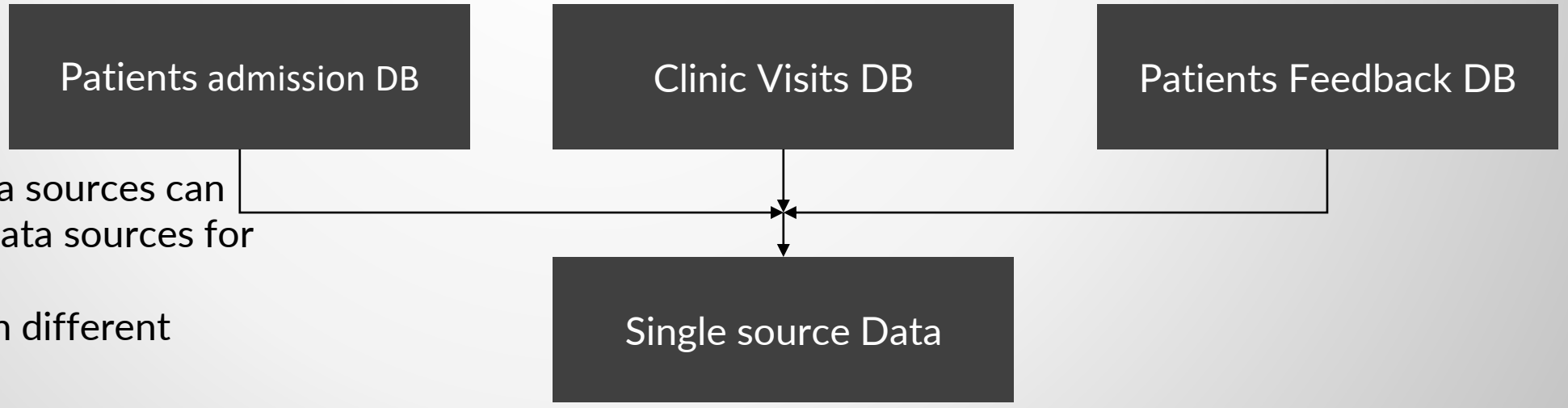
13 Jan 2024

Tool and Concepts

- **Python** with Pandas, NumPy, and scikit-learn for data cleaning, analysis and modeling. It provides data structures and functions ease in manipulation and analysis of large data.
- **SQL** for data integration and manipulation. It will be used mostly be Anderson clinic center to perform querying and managing relational databases. It is best tool to perform data mining.
- **Tableau or Power BI** for data visualization. They will be employed to create visually appealing and interactive dashboards, when presenting findings to management.
- **R** for Statistical analysis, data visualization and data manipulation.
- **Microsoft Excel** will be used to perform basic data cleaning, manipulation and analysis of small datasets.

Data Sources and Preparation requirements

- For Anderson Cancer Center, we will need data from different sources to complete our analysis. The data include:
  - Patient Admission Data: For understanding underlying health conditions.
  - Clinic Visits: For monitoring patient interactions and any required follow-ups.
  - Patient Feedback: To take into consideration on feedback received from patients ensuring satisfaction.
- **Issues to Address:** Data set errors, missing values, inconsistent data, and different formatting in the datasets.



Data from different data sources can be combined into one data sources for the purpose of analysis.
This is achieved through different levels of Normalization

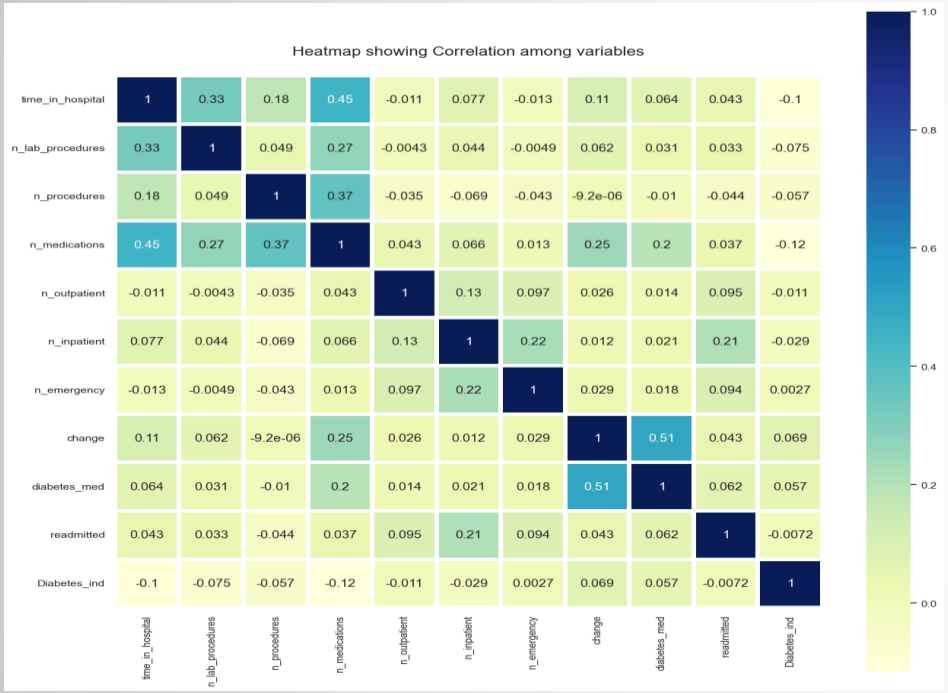
Data Cleaning Techniques and Best Practices

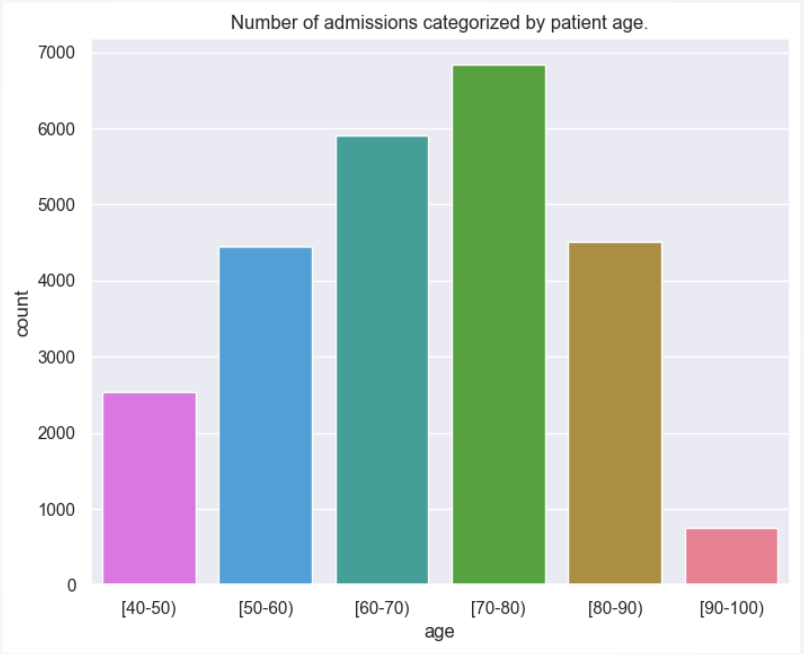
This is the first step of identifying and addressing issues with the data that can negatively impact the accuracy and effectiveness of the analysis or model.

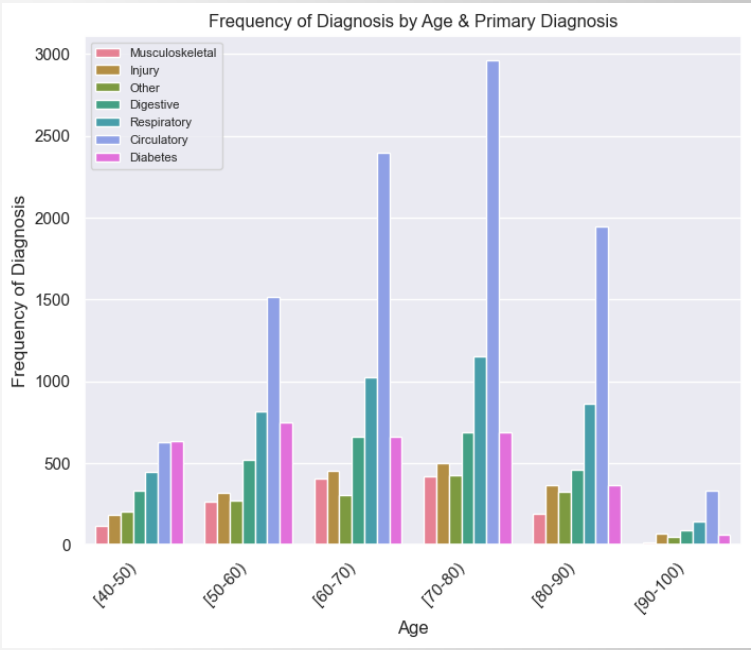
- **Handle Missing Values:** removal or filling missing values in a data set
- **Address Inconsistencies:** Makes data consistent by removing inconsistencies
- **Outlier Detection and Handling:** Identify and handle outliers in patient admission and clinic visit data.
- **Data Validation:** These checks ensures data accuracy and integrity
- **Duplicate Removal:** high level of accuracy and integrity can be obtained when duplicates are removed
- **Identifying and correcting errors**: This can include correcting spelling mistakes, formatting errors, and any other type of mistake present in data.
- **Formatting data:** this makes the data to be coherent when performing analysis.

Data Cleaning: Exploratory Data Analysis

- **Below is a real life analysis of 10 year data of patients readmitted after being discharged obtained from Kaggle.**
- **Upon performing exploratory data analysis, below are the findings.**







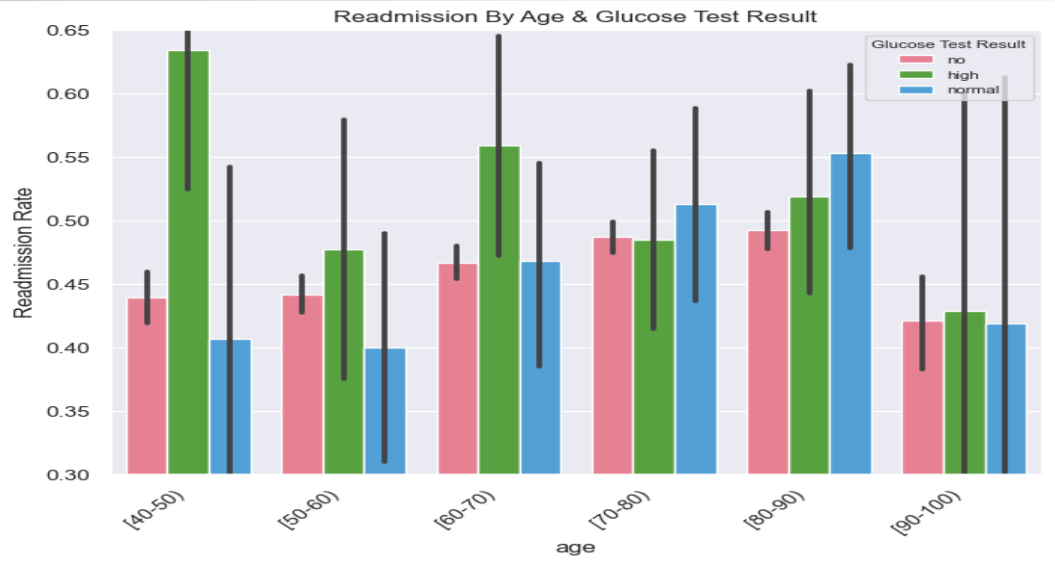
Use of heatmap shows a weak correlation between number of inpatient and readmitted. Based on colour coding, it shows there is a weak correlation between inpatients and likelihood of being readmitted

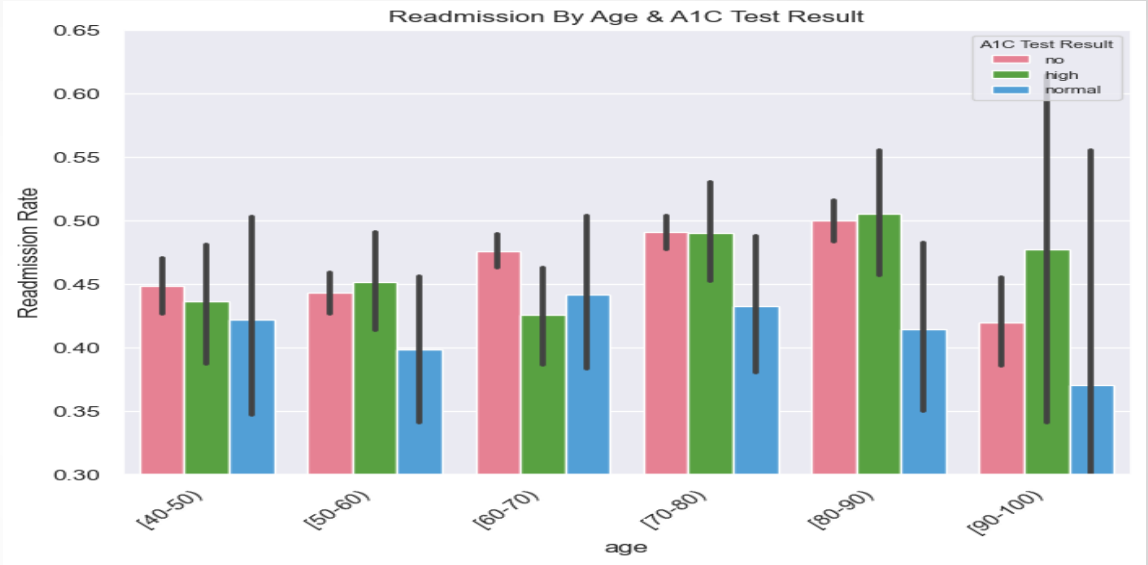
Using the bar chart above to analyze the number of admission per age group, it shows that persons of ages 70-80 are likely to be admitted.

This shows frequency of primary diagnosis, where circulatory diagnosis is the most frequent diagnosis in the hospital facility.

Data Cleaning: Exploratory Data Analysis

- **Exploring the effects of Diabetes Diagnosis on Readmission rates**

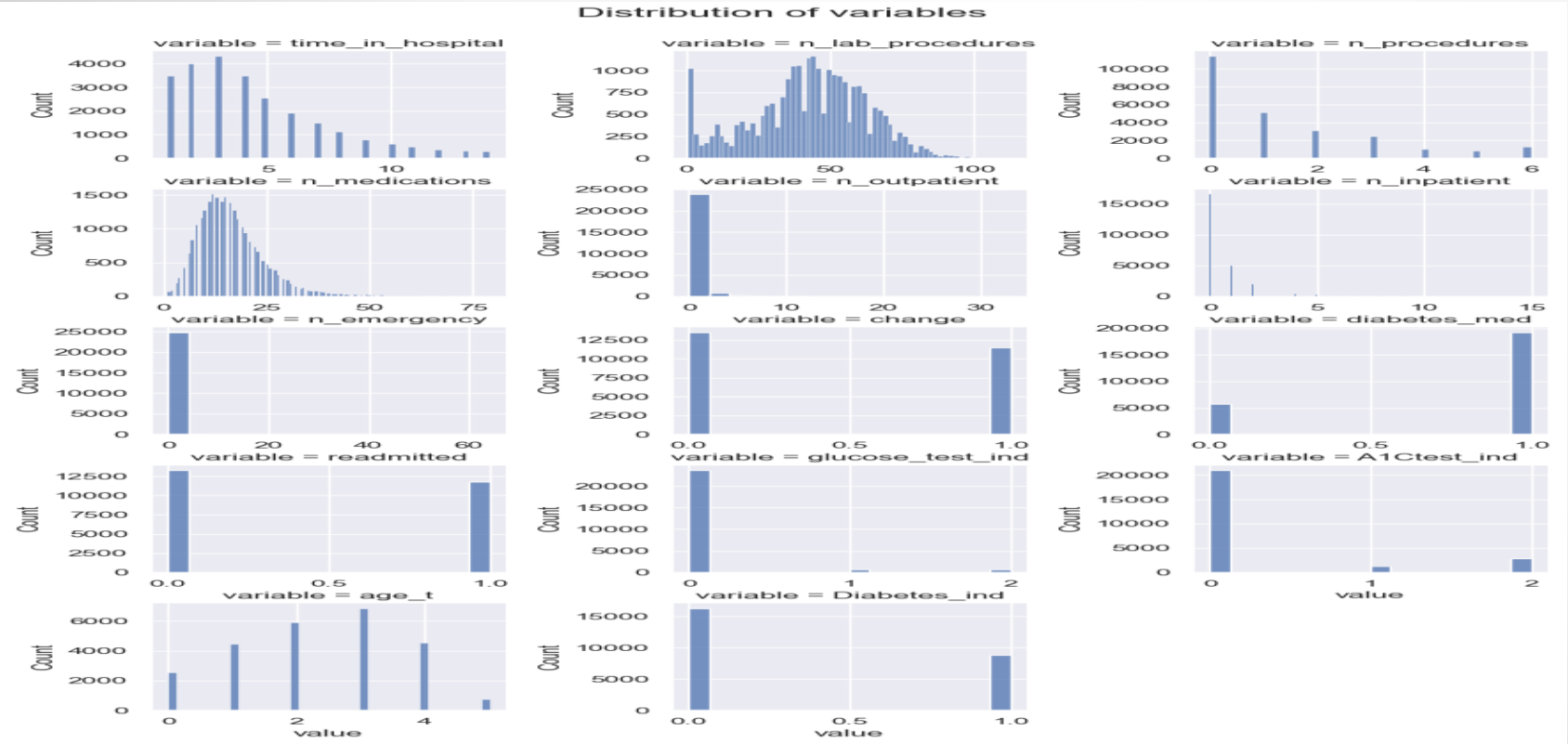




•There is a positive correlation between the number of laboratory procedures conducted throughout a hospital stay and the likelihood of readmission, specifically in the case of patients prescribed diabetes medication. This is seen for the patients who have a high glucose test results.

Data Transformation Methods

- **Normalization:** Standardization of numerical values in a dataset to a common scale.
- **Categorical Variable Encoding:** Converts categorical variables into numerical format for modeling.
- **Feature Engineering:** Create relevant features such as readmission history or patient demographics.
- **Data Reshaping**: in pandas library we use pivot, melt, or reshape data to meet any analytical requirements.



- This is the findings in the data used to show if there is normality or non normality in the data used,
- Once that is established, different normalization techniques can be employed to ensure that the data used is coherent and falls on a normal scale.

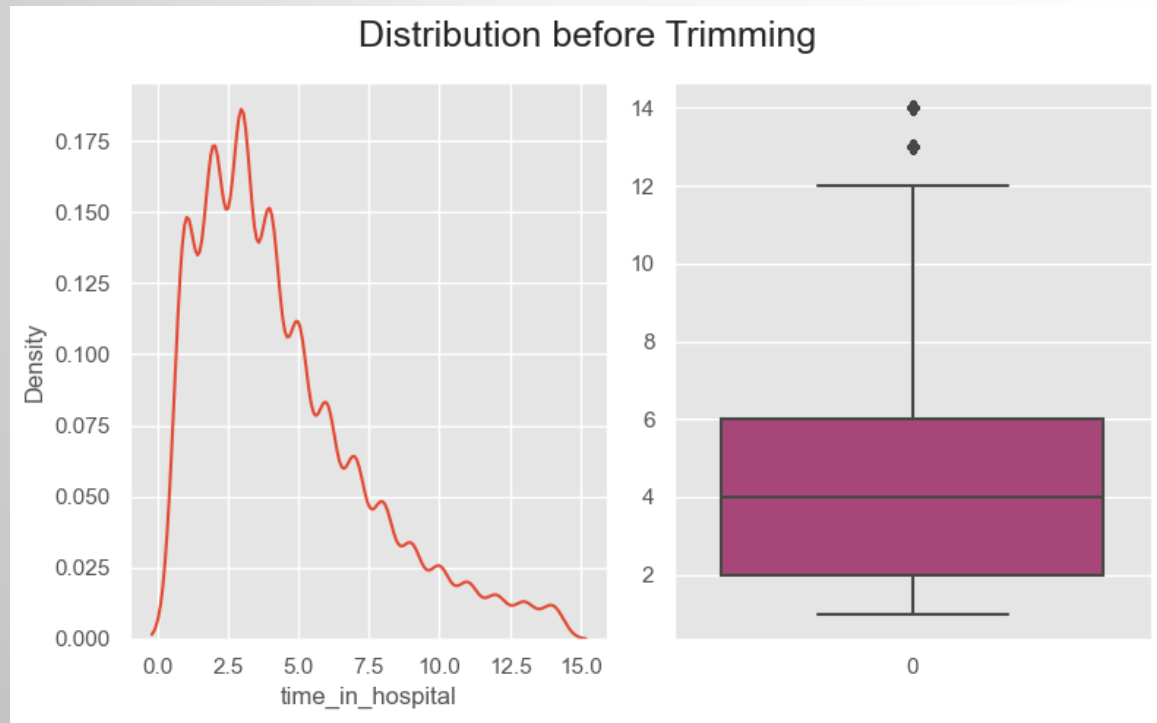
Data Integration Techniques

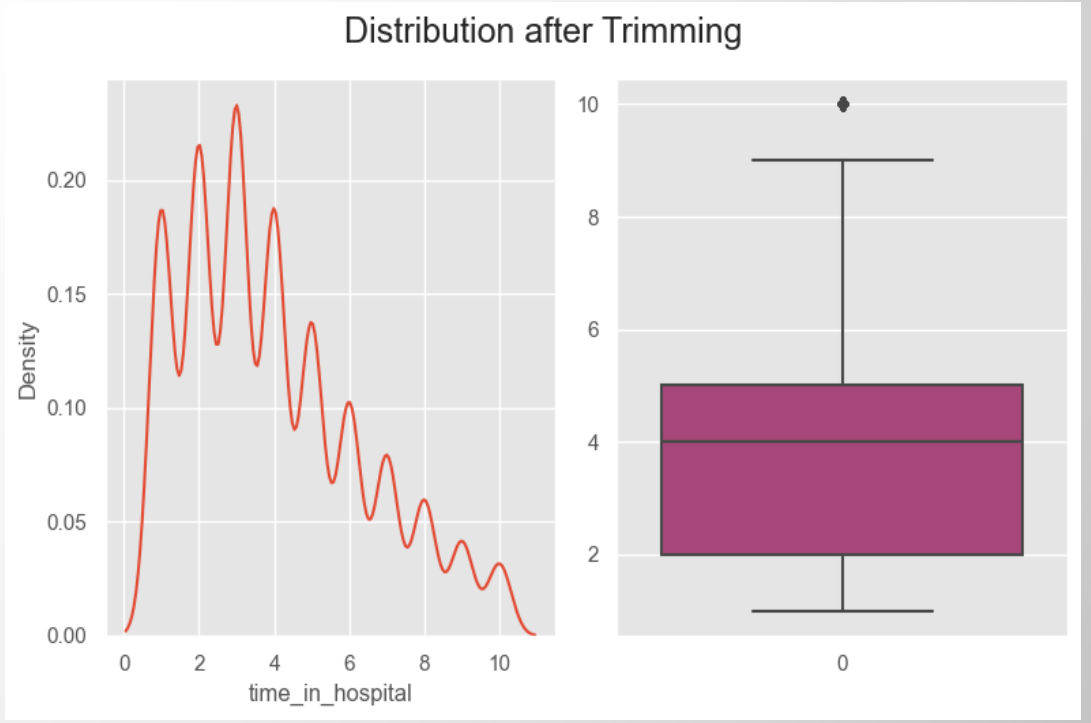
- This involves combining data from multiple sources and combining it into a single source.
- For this analysis, it involves combining data using Microsoft SQL server to create a single document that can be used for the analysis.

- **Specific techniques that can be employed include:**
  - **Merge and Join Operations:** Combine data from different sources using patient common identifying keys.
  - **Data Concatenation:** Used to combine datasets to any desired formats
  - **Master Data Management:** Establishing a master data repository to manage patient records in a consistent manner.

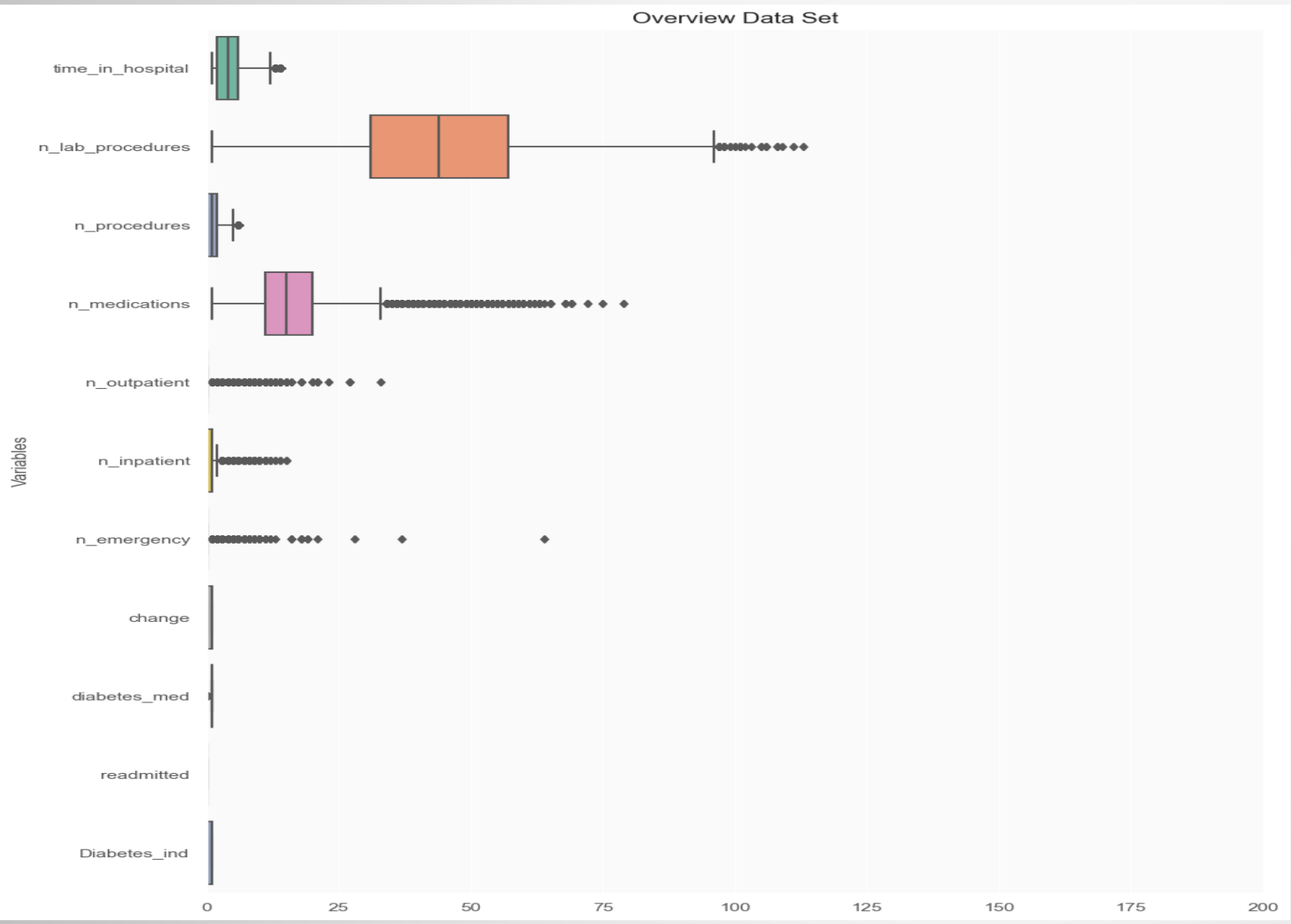
Data Quality and Relevance for Business Analytics Projects

- **Relevance:** Ensure the selected data features align with the goals of predicting patient readmission rates and improving patient care.
- **Consistency:** Standardize data formats and units to maintain consistency across datasets.
- **Accuracy:** Implement validation checks and data cleaning procedures to enhance data accuracy.
- **Timeliness:** Use the most recent and relevant data for analysis.
- **Completeness**: This addresses missing values and filling gaps in the dataset used.





Data Quality and Relevance for Business Analytics Projects



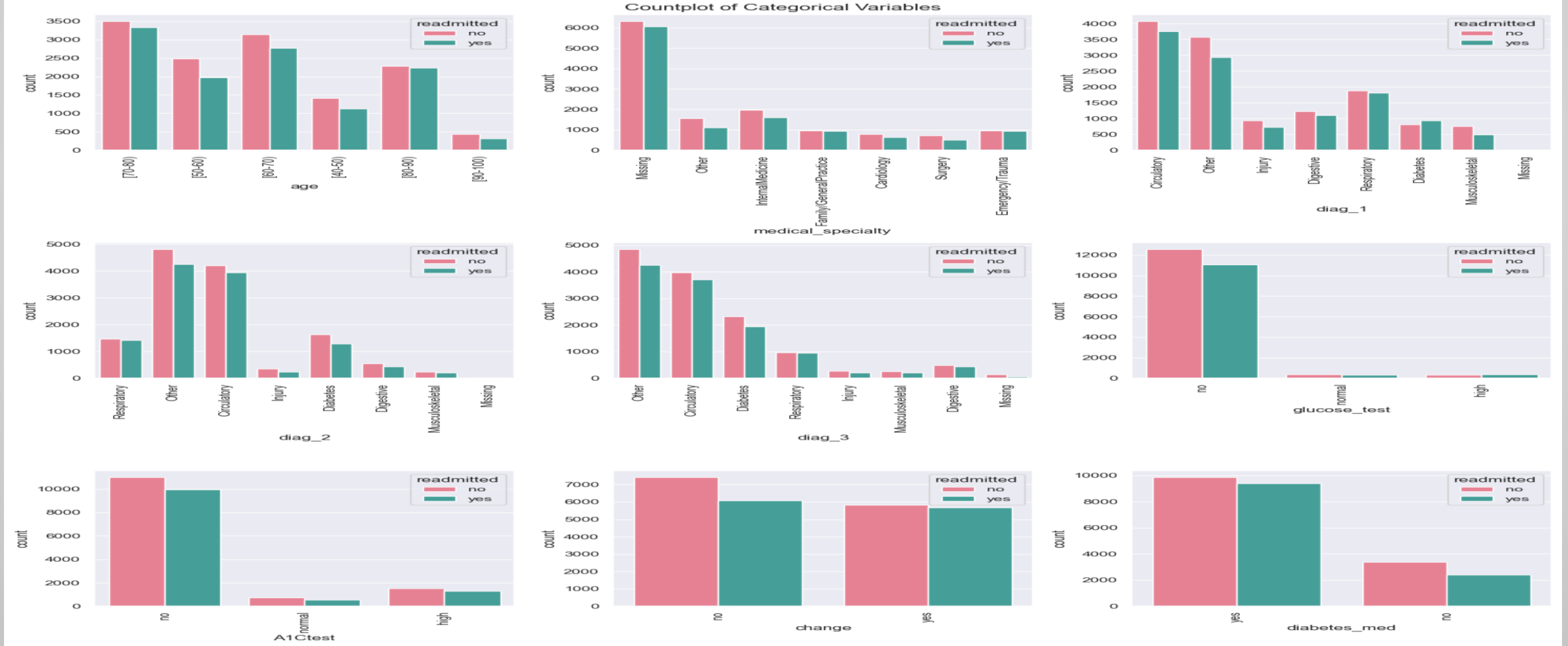
- To ensure accuracy and quality of the data used, outliers can be used to spot the points that are different from the rest of observations.
- These can affect accuracy and reliability of the analysis done.
- Upon analysis of the datasets from the hospital records, various categories of data requires transformation and removal of outliers.

Using Business Analytics Concept(s)

- **Descriptive Analytics:** To conduct analytics to understand what has happened in the past, example in real time data from the medical records, analytics can reveal what happened I the hospital over the past
- **Diagnostic Analysis.** This goes beyond what it happened and answers the question why it happened, example why is it that patients between certain age groups had a high chances of being readmitted.
- **Time Series Analysis:** Identify patterns in patient readmission rates over time, enabling proactive measures.
- **Prescriptive analytics:** Describes actions that needs to be taken to answer to meet desired outcome.
- **Predictive Analytics:** Develop models to predict readmission risks and recommend personalized interventions.

Using Business Analytics Concept(s):

- **Using python pandas library, below can be analyzed from the hospital 10 year records, to reveal specific attributes of different variables in the hospital and make informed decisions accordingly.**



Expected Challenges

- **Data Integration Challenges:** Merging diverse datasets with varying structures.
- **Data Quality Issues:** Addressing errors and inconsistencies in the data.
- **Privacy and Security Concerns**: Ensuring patient data confidentiality.
- **Resource Constraints:** Adequate resources for data preparation and analysis

References

Dubrave. (2023). Predicting Hospital Readmissions. Kaggle https://www.kaggle.com/datasets/dubradave/hospital-readmissions/code