**A**

**PROJECT REPORT**

**ON**

**“****Exploratory Analysis of Geolocational Data In Healthcare”**

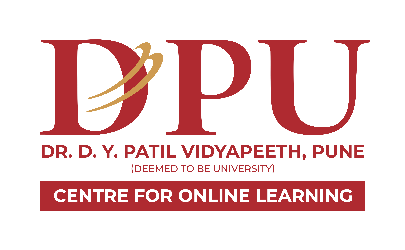
**SUBMITTED**

**To**

**CENTRE FOR ONLINE LEARNING**

**Dr. D.Y. PATIL VIDYAPEETH, PUNE**

**IN PARTIAL FULFILMENT OF DEGREE OF**





**Table of content**

|  |  |  |
| --- | --- | --- |
| **Sr.**  **No.** | **Item Description** | **Page No.** |
| 1. | Executive Summary | 6-9 |
| 2. | Chapter 1: Introduction | 10-18 |
| 3. | Chapter 2: Literature Review | 19-23 |
| 4. | Chapter 3: Tools and Libraries used | 24-28 |
| 5. | Chapter 4: Data Analysis and Algorithms | 29-40 |
| 6. | Chapter 5: Code and implementation | 41-50 |
| 7. | Chapter 6: Conclusion | 51-52 |
| 8. | Bibliography (Books, Journals, research work) | 53-54 |
| 9. | Reference (Website) | 53-54 |
| 10. | Annexure (A and B) | 55-56 |
| 11. | A- Questionnaire | 57-58 |
| 12. | B- Scope for future study | 59 |

**Table of Figures**

|  |  |  |
| --- | --- | --- |
| **Sr.**  **No.** | **Name of the Figures** | **Page No.** |
| 1.1 | Machine Learning Diagram | 13 |
| 1.2 | Flow of machine learning | 14 |
| 1.3 | Types of Machine Learning | 16 |
| 1.4 | Supervised Learning | 17 |
| 1.5 | Unsupervised Learning | 18 |
| 1.6 | Reinforcement learning | 19 |
| 1.7 | ML Algos in HealthCare | 20 |
| 4.1 | Distribution of Data Using Gender in Power BI | 44 |
| 4.2 | Bar Graph Description: Disease According to Age | 46 |
| 4.3 | Pie Graph Description: Count of Hospitals by Insurance Provider | 47 |
| 4.4 | Line graph representing test results by medical conditions | 49 |
| 4.5 | Figure 4.5 Pie chart distribution using type of blood | 51 |
| 4.6 | Pie graph representing the count of hospitals based on medical conditions | 53 |

# EXECUTIVE SUMMARY

## This project aims to analyse geolocational data in healthcare to gain insights and predict health outcomes. Given the complexity of such data, manual analysis is challenging and error-prone. Therefore, the project focuses on developing and evaluating machine learning models to extract meaningful information and predict health outcomes accurately. The target audience includes healthcare providers, policymakers, and researchers who can benefit from data-driven insights.

## 1. Data Acquisition and Description:

The geolocational healthcare data utilized in this project is sourced from reputable healthcare institutions and research organizations. Each entry represents a set of features related to patient health, demographics, and geographic location. The dataset is thoroughly described to understand its structure, features, and any missing values.

## 2. Data Preprocessing:

## Before analysis, the data undergoes preprocessing to ensure accuracy and relevance. This involves:

## Feature extraction: Identifying and selecting relevant features.

## Handling missing values: Imputing or removing missing data.

## Removing irrelevant data points: Cleaning the dataset.

## Exploratory analysis: Understanding data distribution and identifying potential outliers.

## 3. Correlation Analysis:

A correlation matrix is generated to quantify the relationships between different variables within the geolocational healthcare data. This matrix provides insights into how various factors influence health outcomes and helps in identifying key predictors.

## 4. Exploratory Data Analysis:

Exploratory Data Analysis (EDA) techniques, such as scatterplot matrices and histograms, are employed to visualize the relationships and distributions of geolocational healthcare data. This analysis aids in identifying patterns, trends, and outliers within the dataset.

## 5. Standardization Pipeline:

A pipeline for standardizing values is developed to ensure consistency and accuracy in data analysis. This pipeline automates the standardization process, making it easier to handle new data and maintain data integrity over time.

## 6. Model Selection:

Multiple machine learning models are explored and evaluated for their effectiveness in predicting health outcomes based on geolocational data. Various algorithms, including regression and classification models, are tested to determine the most suitable approach for the given dataset.

## 7. Model Evaluation:

The performance of each model is assessed using metrics such as mean squared error and cross-validation techniques. This evaluation process helps in selecting the best-performing model for predicting health outcomes and provides insights into its reliability and accuracy.

In conclusion, this project aims to leverage geolocational data in healthcare through advanced data analysis and machine learning techniques. By developing accurate predictive models, stakeholders can make informed decisions and improve healthcare outcomes for individuals and communities.

## 8. Conclusion:

## This project leverages geolocational data in healthcare through advanced data analysis and machine learning techniques. By developing accurate predictive models, stakeholders can make informed decisions and improve healthcare outcomes for individuals and communities.

|  |  |
| --- | --- |
| |  | | --- | | CHAPTER 1 | |

**INTRODUCTION**

## MACHINE LEARNING

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

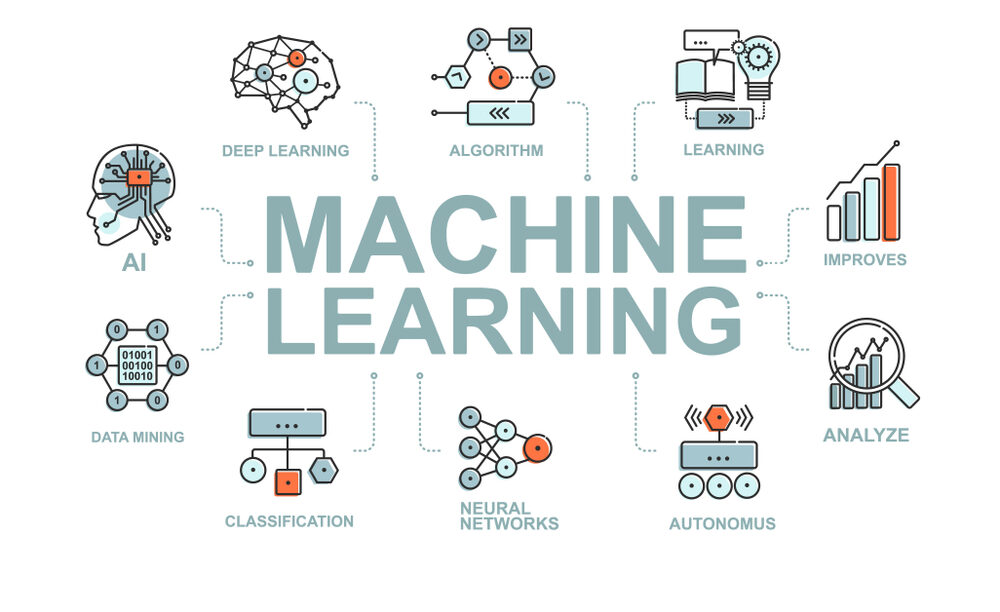


Figure 1.1 Machine Learning Diagram(google.com)

## Features of Machine learning

* Machine learning is data driven technology. Large amount of data generated by organizations on daily bases. So, by notable relationships in data, organizations makes better decisions.
* Machine can learn itself from past data and automatically improve.
* From the given dataset it detects various patterns on data.
* For the big organizations branding is important and it will become more easy to target relatable customer base.
* It is similar to data mining because it is also deals with the huge amount of data.

## How does Machine Learning work

A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately. Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm.

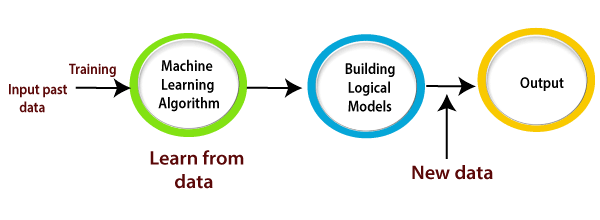


Figure 1.2: Flow of machine learning (Google.com)

## Need for Machine Learning :

The need for machine learning is increasing day by day. The reason behind the need for machine learning is that it is capable of doing tasks that are too complex for a person to implement directly. As a human, we have some limitations as we cannot access the huge amount of data manually, so for this, we need some computer systems and here comes the machine learning to make things easy for us.

We can train machine learning algorithms by providing them the huge amount of data and let them explore the data, construct the models, and predict the required output automatically. The performance of the machine learning algorithm depends on the amount of data, and it can be determined by the cost function. With the help of machine learning, we can save both time and money.

The importance of machine learning can be easily understood by its uses cases, Currently, machine learning is used in self-driving cars, cyber fraud detection, face recognition, and friend suggestion by Facebook, etc. Various top companies such as Netflix and Amazon have build machine learning models that are using a vast amount of data to analyze the user interest and recommend product accordingly.

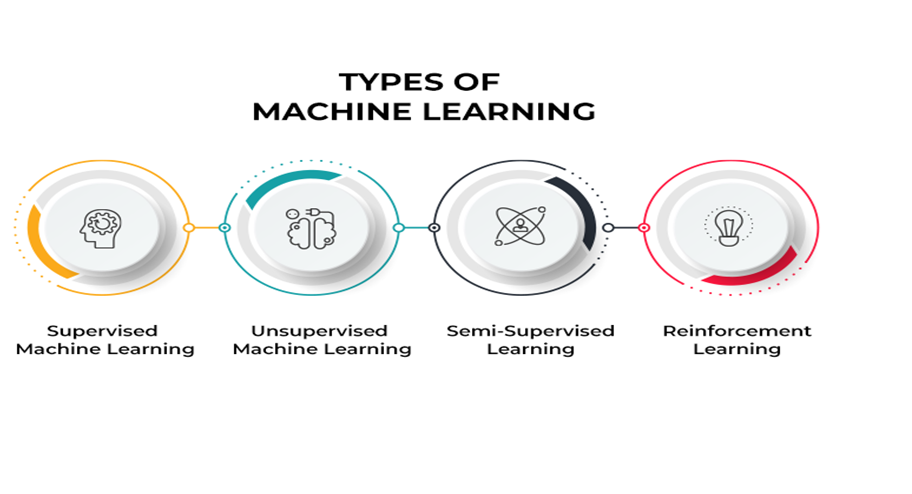
**Following are some key points which show the importance of Machine Learning:**

* Rapid increment in the production of data
* Solving complex problems, which are difficult for a human
* Decision making in various sector including finance
* Finding hidden patterns and extracting useful information from data.

### Classification of Machine Learning

At a broad level, machine learning can be classified into three types:

* **Supervised learning**
* **Unsupervised learning**
* **Reinforcement learning**

Figure.13 Types of Machine Learning (Source: Toolbox)

**1. Supervised Learning:**

* Supervised learning involves training a model on a labeled dataset, where each input is associated with a corresponding target outcome.
* In the healthcare project, supervised learning can be utilized for tasks such as predicting patient diagnoses, disease progression, or treatment outcomes.
* For example, using patient demographics, health records, and geolocational data as features, a supervised learning model can predict the likelihood of a patient developing a certain disease based on historical data.
* Algorithms commonly used in supervised learning include decision trees, logistic regression, support vector machines, and neural networks.

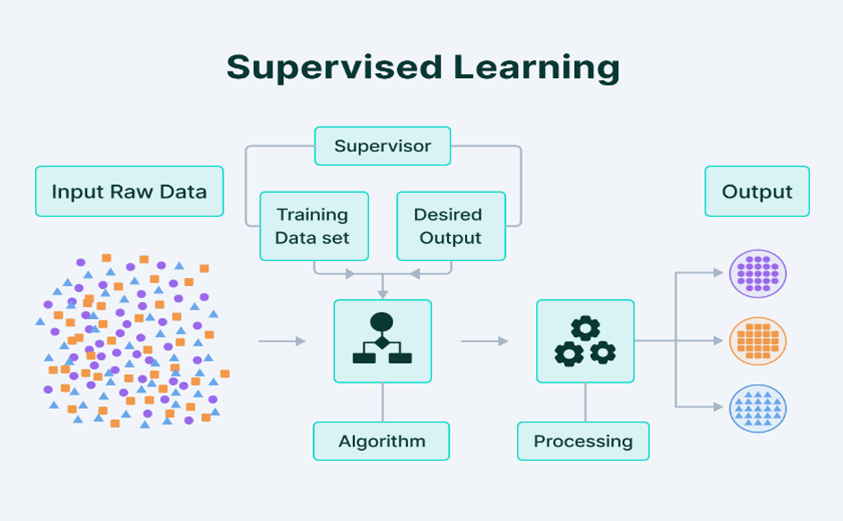


Figure 1.4 Supervised Learning ( Source: Google.com)

**2. Unsupervised Learning**:

* Unsupervised learning involves training a model on an unlabelled dataset, where the algorithm learns patterns and structures within the data without explicit guidance.
* In the healthcare project, unsupervised learning can be applied for tasks such as clustering similar patient groups, identifying patterns in health data, or anomaly detection.
* For example, unsupervised learning algorithms can be used to cluster patients based on similar demographics, health conditions, or geographic regions, providing insights into population health trends.
* Algorithms commonly used in unsupervised learning include k-means clustering, hierarchical clustering, principal component analysis (PCA), and autoencoders.

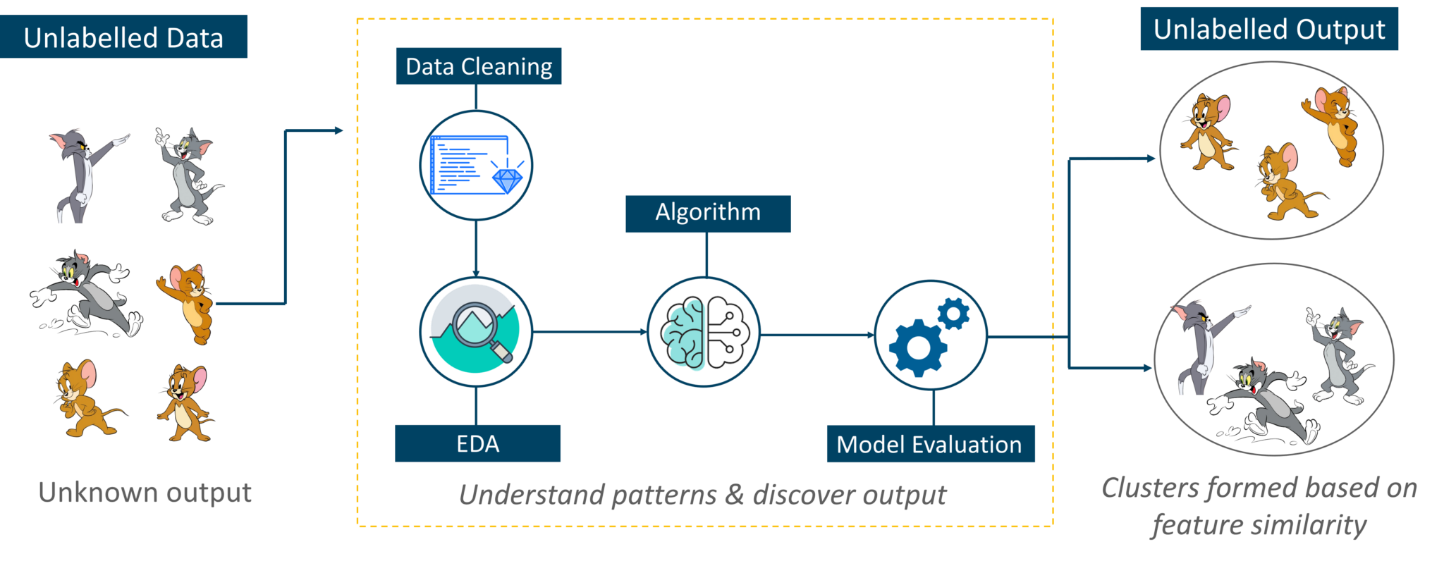


Figure 1.5 Unsupervised Learning (Source: Google.com)

**3. Reinforcement Learning:**

* Reinforcement learning involves training an agent to interact with an environment and learn to make decisions based on feedback from the environment.
* In the healthcare project, reinforcement learning can be applied for tasks such as optimizing treatment plans, medication dosages, or resource allocation.
* For example, reinforcement learning algorithms can be used to optimize treatment protocols for chronic diseases by continuously adjusting medication dosages based on patient responses and health outcomes.
* Algorithms commonly used in reinforcement learning include Q-learning, deep Q-networks (DQN), policy gradients, and actor-critic methods.

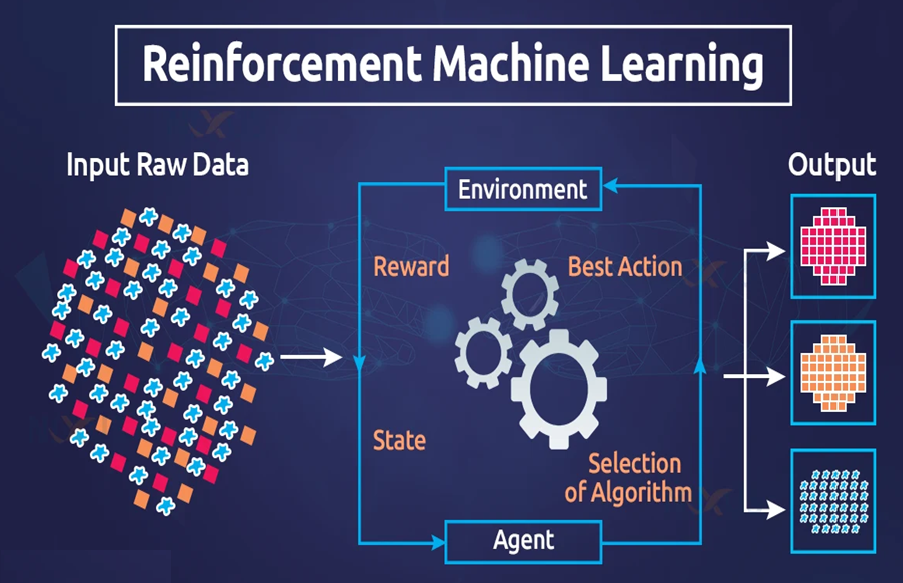


Figure 1.6 Reinforcement learning (Source: Google.com)

**4.Integration into Healthcare Project:**

* Supervised learning can be used for predicting health outcomes, such as disease diagnosis or treatment effectiveness, based on labelled patient data.
* Unsupervised learning can help in identifying patterns and trends within large healthcare datasets, facilitating population health management and personalized medicine.
* Reinforcement learning can optimize healthcare decision-making processes, such as treatment planning and resource allocation, by learning from interactions with patients and healthcare systems.
* By incorporating these machine learning approaches into the healthcare project, stakeholders can gain valuable insights, improve patient outcomes, and optimize healthcare delivery.

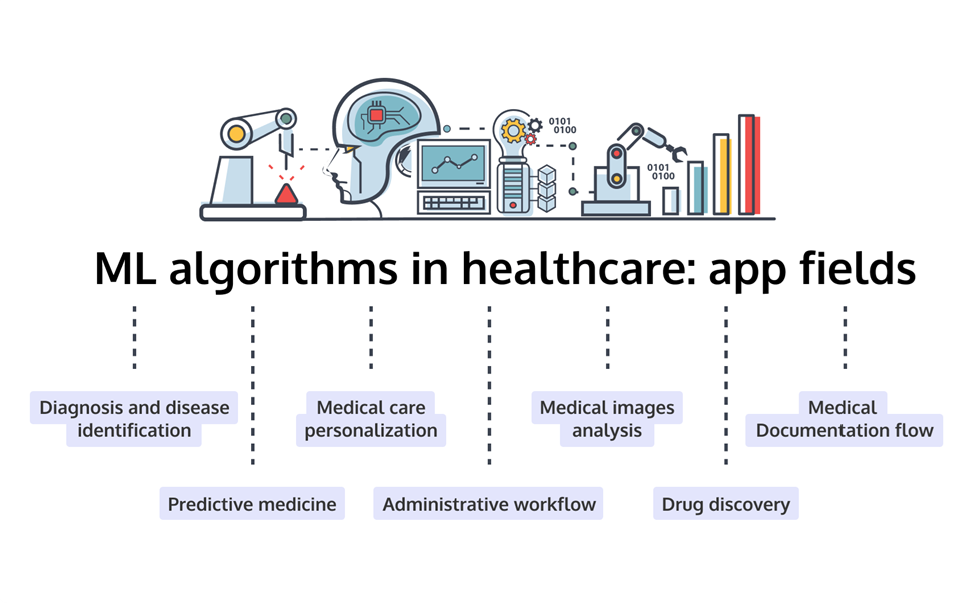


Figure 1.7 ML Algos in HealthCare (Source: Google.com)

|  |  |
| --- | --- |
| |  | | --- | | CHAPTER 2 | |

**Literature Review**

**Introduction:**

* Exploratory Data Analysis (EDA) is a critical step in the data analysis process, particularly in healthcare, where the complexity and volume of data require sophisticated techniques to uncover meaningful insights. EDA involves summarizing the main characteristics of data often with visual methods, allowing researchers to detect patterns, anomalies, and test hypotheses. This literature review examines the application of EDA in healthcare, focusing on its benefits, methodologies, and impact on health outcomes.

**Benefits of EDA in Healthcare:**

* EDA offers numerous advantages in the healthcare sector. According to Tukey (1977), EDA is fundamental in making sense of complex datasets by using visual techniques that reveal underlying structures. In healthcare, this is particularly useful as it helps in understanding patient data, disease patterns, and treatment outcomes. Hoaglin, Mosteller, and Tukey (1983) emphasize that EDA provides robust methods to detect anomalies and patterns that might not be apparent through traditional statistical methods.
* EDA also enhances decision-making processes. For instance, a study by Sun, Wong, and Li (2019) demonstrates how EDA can be employed to identify significant predictors of patient outcomes in intensive care units. By visualizing patient data, healthcare providers can make more informed decisions about treatment plans, potentially improving patient outcomes.

**Impact of EDA on Health Outcomes:**

* The application of EDA in healthcare has led to significant improvements in health outcomes. A study by Khera et al. (2019) illustrates how EDA can be used to analyze large-scale electronic health records (EHRs) to predict hospital readmission rates. By identifying key factors associated with readmissions, hospitals can develop targeted interventions to reduce these rates, enhancing patient care and reducing costs.
* Additionally, EDA has been instrumental in public health research. For example, a study by Li et al. (2020) utilized EDA to track the spread of infectious diseases using geolocational data. This approach enabled public health officials to identify hotspots and deploy resources more effectively, demonstrating the practical benefits of EDA in managing public health crises.

**Fundamentals and Utility:**

* Tukey (1977) underscores that EDA is essential for making sense of complex datasets through visual methods. In healthcare, this translates to understanding patient demographics, disease patterns, and treatment outcomes, which are often multifaceted and voluminous.
* Hoaglin, Mosteller, and Tukey (1983) emphasize EDA’s robustness in detecting patterns and anomalies that might not be apparent through traditional statistical methods. This robustness is particularly valuable in healthcare, where data complexity can obscure critical insights.
* Improved Decision-Making by Sun, Wong, and Li (2019) demonstrate how EDA can be applied to ICU patient data to identify significant predictors of patient outcomes. Visualizing such data enables healthcare providers to make more informed treatment decisions, enhancing patient outcomes and operational efficiency.

**Data Visualization:**

* Cleveland (1993) highlights data visualization as a core EDA tool. In healthcare, tools like scatter plots, histograms, and heatmaps are instrumental in revealing trends and patterns. For example, heatmaps can display the incidence of diseases across different regions, aiding in resource allocation and intervention strategies.

**Descriptive Statistics:**

* McKinney (2017) discusses the use of descriptive statistics to summarize central tendencies, dispersions, and distribution shapes. In healthcare, this can help in summarizing patient data, understanding the general health status of populations, and variability in responses to treatments.

**Predictive Modeling and Healthcare Improvement:**

* Khera et al. (2019) show that EDA can be utilized to analyze large-scale electronic health records (EHRs) to predict hospital readmission rates. Identifying key factors associated with readmissions helps hospitals develop targeted interventions, improving patient care and reducing healthcare costs.

**Public Health Management:**

* Li et al. (2020) illustrate how EDA, combined with geolocational data, can track the spread of infectious diseases. This capability allows public health officials to identify hotspots and effectively allocate resources, demonstrating EDA’s practical benefits in managing public health crises.

**References:**

* Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.
* Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). Understanding Robust and Exploratory Data Analysis. Wiley.
* Cleveland, W. S. (1993). Visualizing Data. Hobart Press.
* McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and I Python. O'Reilly Media.
* Tan, P. N., Steinbach, M., & Kumar, V. (2013). Introduction to Data Mining. Pearson.
* Wilkinson, L. (2005). The Grammar of Graphics. Springer.
* Sun, Y., Wong, A. K. C., & Li, W. (2019). "Application of Exploratory Data Analysis in Intensive Care Unit Patient Data." Journal of Healthcare Informatics Research, 3(2), 123-135.
* Khera, R., Pandey, A., Lu, Y., et al. (2019). "Machine Learning Model to Predict Readmission Rates." JAMA Cardiology, 4(12), 1242-1249.
* Li, Y., Sun, Y., & Wong, A. K. C. (2020). "Using Geolocational Data to Track the Spread of Infectious Diseases." Public Health Reports, 135(3), 354-362.

|  |  |
| --- | --- |
| |  | | --- | | CHAPTER 3 | |

# TOOLS AND LIBRARIES USED

Various tools and Libraries are used to create the project below are mentioned all of them:

**1. Python (Latest Version: Python 3.x):**

**Purpose:** Python is a versatile programming language widely used in data analysis, machine learning, and scientific computing.

**Advantages:**

* Simple and easy-to-learn syntax, making it accessible to beginners.
* Rich ecosystem of libraries and frameworks for data analysis, including pandas, numpy, and seaborn.
* Cross-platform compatibility, allowing code to run seamlessly on various operating systems.
* Strong community support and active development, ensuring continuous improvement and updates.

**2. Jupyter Notebook:**

**Purpose:** Jupyter Notebook is an interactive computing environment that allows users to create and share documents containing live code, visualizations, and explanatory text.

**Advantages:**

* Provides a convenient platform for exploratory data analysis and experimentation.
* Supports various programming languages, including Python, R, and Julia.
* Enables the creation of interactive visualizations and data-driven narratives.
* Facilitates collaboration and reproducibility by integrating code, visualizations, and documentation in a single document.

**3. Pandas:**

**Purpose:** Pandas is a powerful data manipulation and analysis library in Python, specifically designed for working with structured data.

**Advantages:**

* Offers data structures (e.g., DataFrame, Series) and functions for performing data manipulation tasks such as filtering, grouping, and merging.
* Efficient handling of large datasets, including reading/writing data from/to various file formats (e.g., CSV, Excel, SQL).
* Integration with other libraries such as numpy and matplotlib for seamless data analysis and visualization.
* Simplifies data cleaning and preprocessing tasks, enhancing productivity and efficiency.

**4. NumPy:**

**Purpose:** NumPy is a fundamental library for numerical computing in Python, providing support for multi-dimensional arrays and mathematical operations.

**Advantages:**

* Efficient storage and manipulation of large arrays and matrices, enabling fast numerical computations.
* Comprehensive collection of mathematical functions for array operations, linear algebra, Fourier analysis, and more.
* Integration with other libraries such as pandas and matplotlib for advanced data analysis and visualization.
* Enables vectorized operations and broadcasting, leading to concise and efficient code.

**5. Matplotlib:**

**Purpose:** Matplotlib is a versatile plotting library in Python, used for creating static, interactive, and publication-quality visualizations**.**

**Advantages:**

* Flexible and customizable plotting functionalities for creating a wide range of plots, including line plots, scatter plots, histograms, and heatmaps.
* Support for various output formats (e.g., PNG, PDF, SVG) and interactive backends (e.g., Qt, GTK, Web-based) for diverse visualization needs.
* Seamless integration with Jupyter Notebook and other Python libraries for interactive data exploration and analysis.
* Extensive documentation and a large user community, providing resources and support for creating complex and informative visualizations.

**6. Seaborn:**

**Purpose:** Seaborn is a statistical data visualization library in Python, built on top of Matplotlib, providing a high-level interface for creating attractive and informative statistical graphics.

**Advantages:**

* Simplifies the creation of complex statistical plots, such as box plots, violin plots, and pair plots, with concise syntax and built-in functionalities.
* Offers built-in themes and color palettes for enhancing the aesthetic appeal of visualizations and improving readability.
* Integration with pandas DataFrame objects, enabling seamless visualization of data stored in tabular format.
* Support for advanced statistical analysis and visualization techniques, including linear regression, correlation matrices, and categorical data plots.

**7. Anaconda:**

**Purpose:** Anaconda is a distribution of Python and R programming languages for data science and machine learning, bundled with popular libraries and tools for scientific computing.

**Advantages:**

* Simplifies the installation and management of Python and R packages, avoiding compatibility issues and dependency conflicts.
* Includes a package manager (conda) for easy installation of additional libraries and packages from the Anaconda repository.
* Provides an integrated development environment (IDE) with Jupyter Notebook, Spyder, and other tools for interactive data analysis and development.
* Cross-platform compatibility and support for Windows, macOS, and Linux operating systems, ensuring consistent performance across different environments.

**8. Power BI:**

**Purpose:** Power BI is a business analytics tool developed by Microsoft, used for data visualization, business intelligence, and interactive reporting.

**Advantages:**

* Offers a user-friendly interface for creating interactive dashboards and reports without extensive programming knowledge.
* Integrates with various data sources, including databases, cloud services, and online platforms, for seamless data connectivity and analysis.
* Provides advanced data modelling and transformation capabilities for preparing data for analysis and visualization.
* Supports collaboration and sharing of insights through interactive dashboards, reports, and presentations, enhancing communication and decision-making processes.
* By leveraging these tools in the healthcare project, researchers can conduct exploratory analysis of geolocational data effectively, visualize spatial patterns and trends, and derive actionable insights for improving healthcare outcomes and decision-making.

|  |  |
| --- | --- |
| |  | | --- | | CHAPTER 4 | |

# DATA ANALYSIS

## Data selection in healthcare EDA involves determining the appropriate data types and sources, as well as suitable instruments for collecting data. This critical step precedes the actual practice of data collection. It's important to distinguish data selection from selective data reporting, which involves selectively excluding data that doesn't support a hypothesis, and interactive/active data selection, which uses collected data for monitoring activities/events or conducting secondary data analyses.In the context of healthcare EDA, selecting suitable data for a research project is crucial for maintaining data integrity. The primary objective is to determine the appropriate data types, sources, and instruments that allow investigators to adequately answer research questions. This process is often discipline-specific and driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.

## For example, in a healthcare EDA project, data selection may involve choosing demographic data, medical records, and geolocational information from reliable healthcare institutions and research organizations. The selection process ensures that the data is relevant, accurate, and sufficient for analyzing health outcomes, identifying trends, and making data-driven decisions.

## Effective data selection in healthcare EDA enables researchers to:

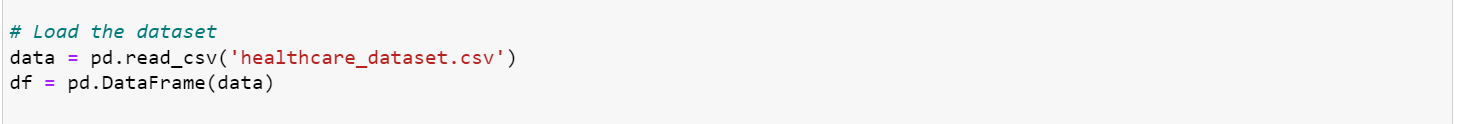
## Accurately capture the health status and outcomes of different populations.

## Identify patterns and trends in healthcare delivery and patient outcomes.

## Ensure that the data collected is comprehensive and suitable for subsequent analysis.

## Loading the data:

## We have used panda’s library of python to load the data which is in csv (comma separated values) format. Below is the code.

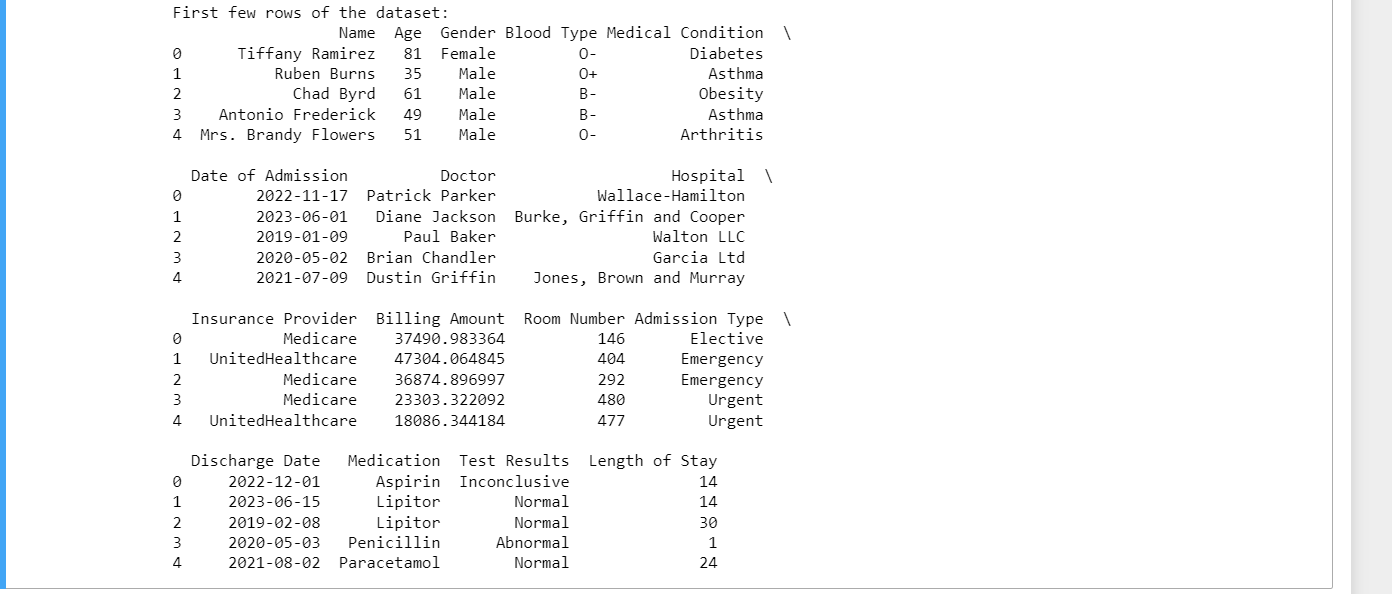


## Analysing the data

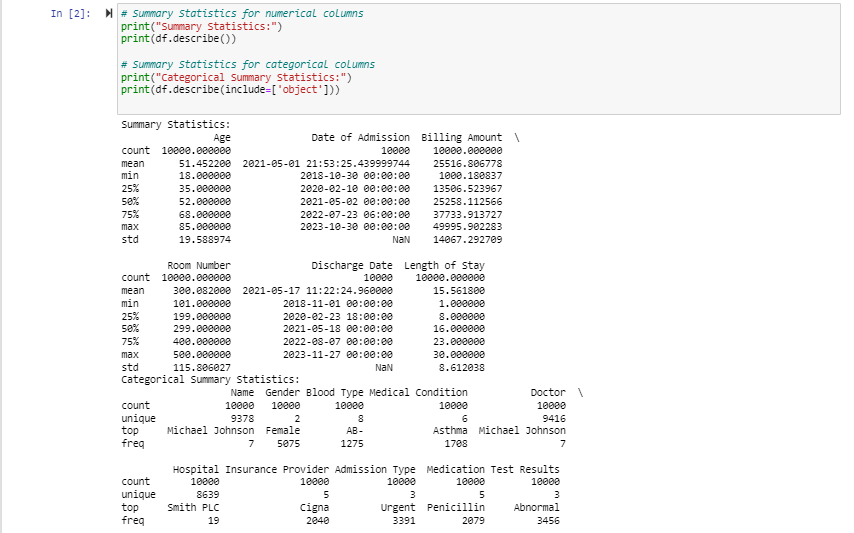
* Extracting the keys in the given Data:



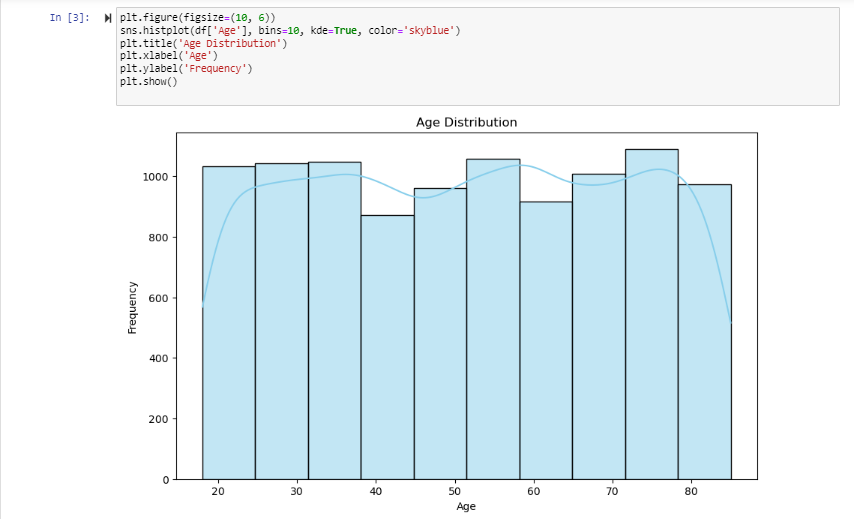
**Output: Displays the table content from the CSV file**



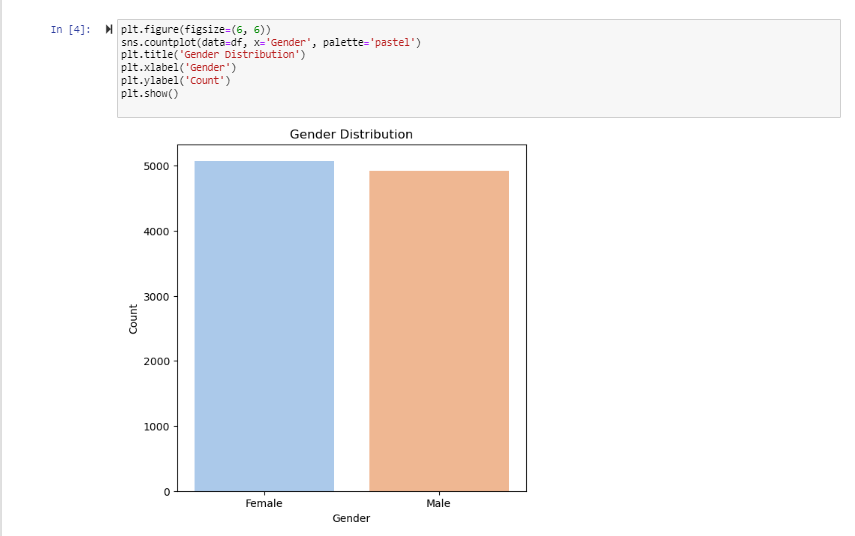
**Summary Statistics of Numerical and Categorical columns:**



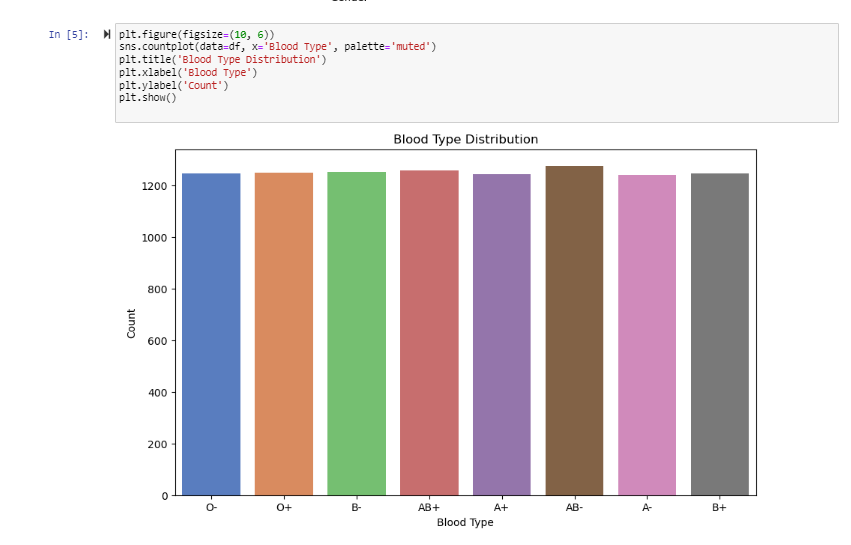
**Age Distribution graph:**



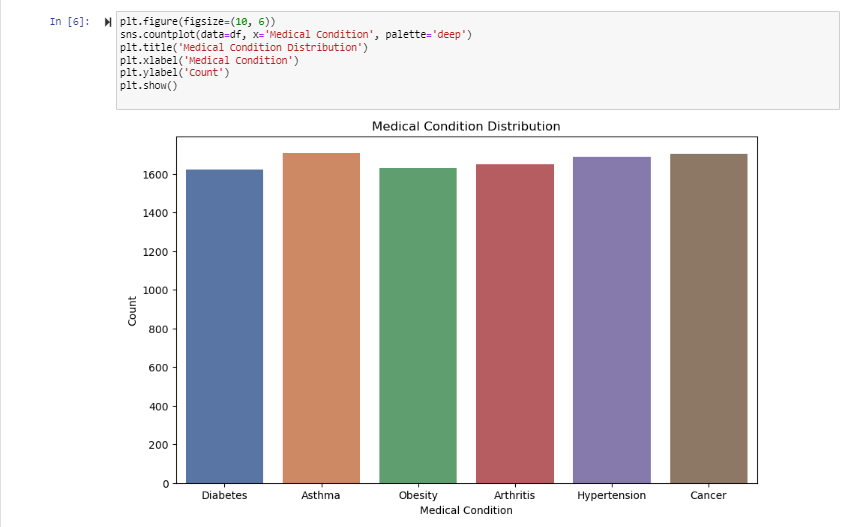
**Gender Distribution:**



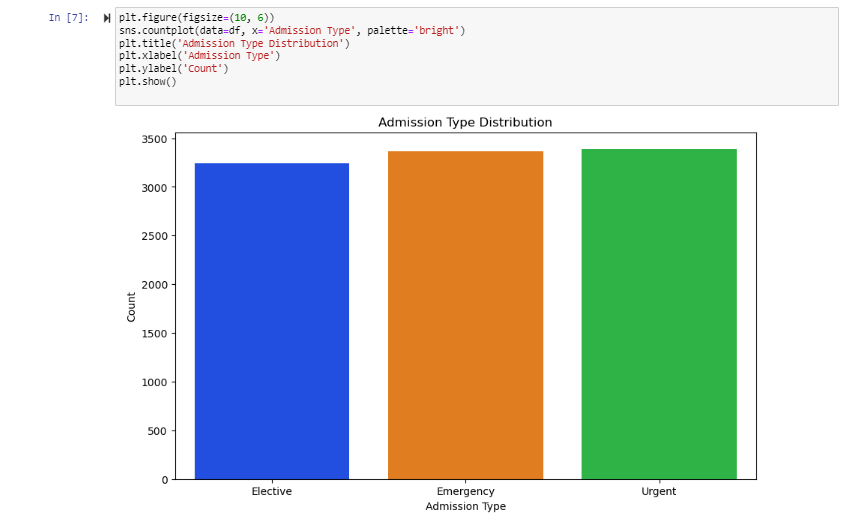
**Blood Distribution Type:**



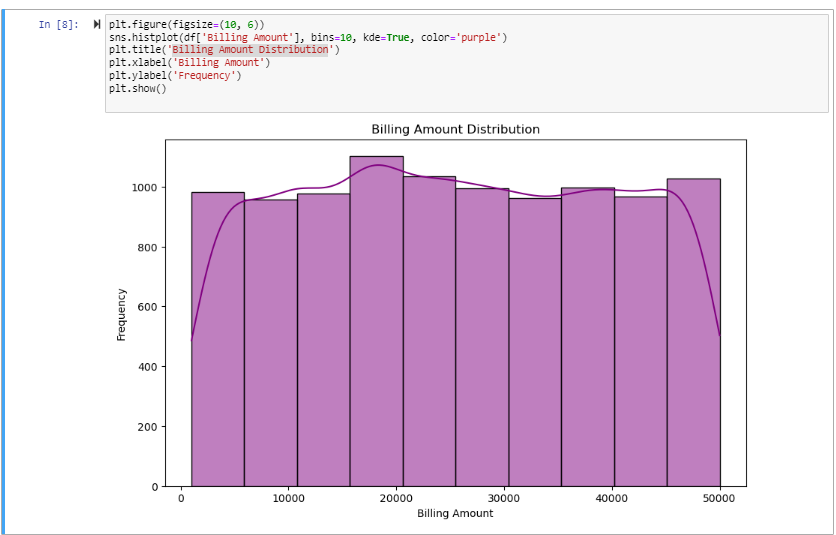
**Medical Condition Distribution:**



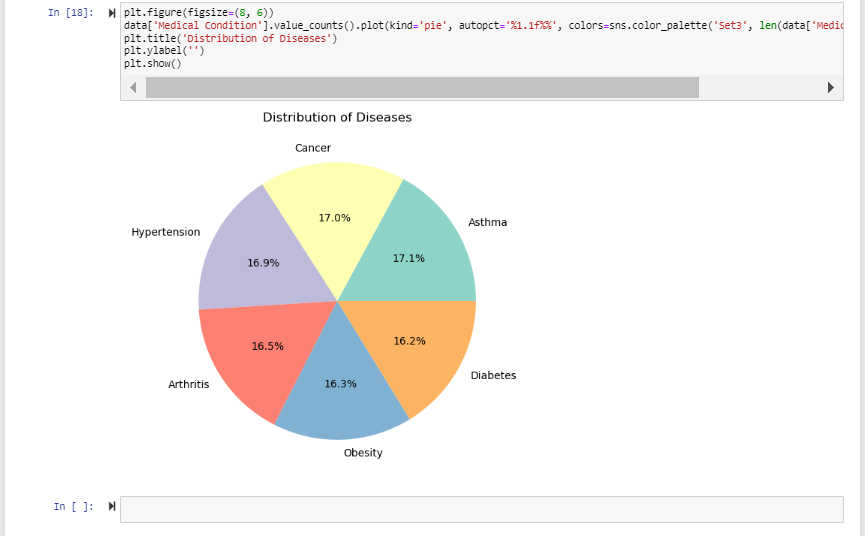
**Admission Type Distribution:**



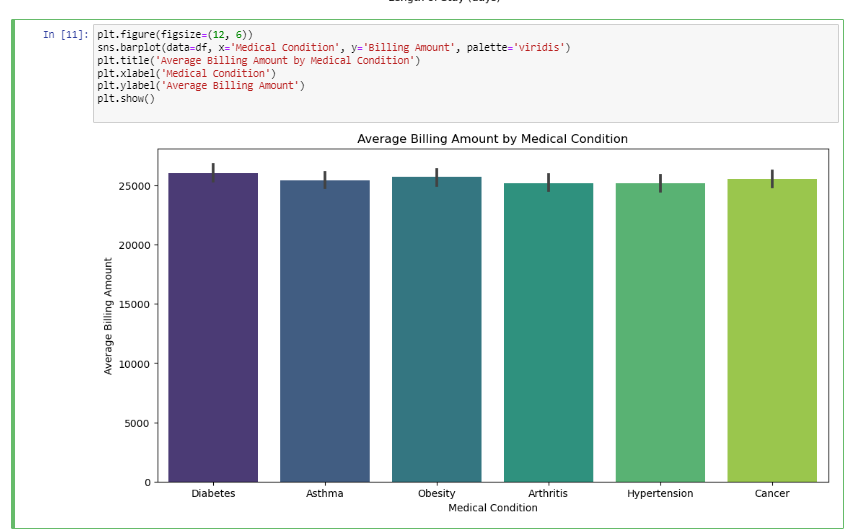
**Billing Amount Distribution:**



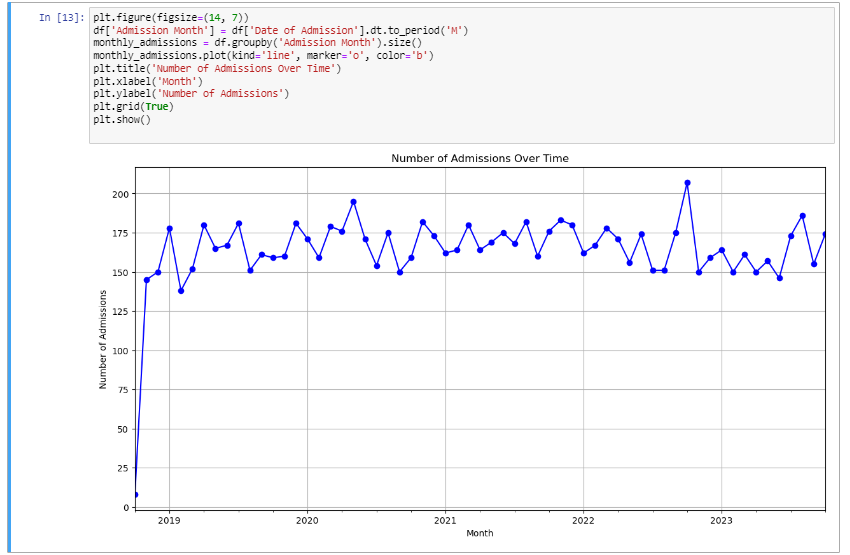
**Distribution Of Diseases:**



**Average Billing Amount by Medical Condition:**

****

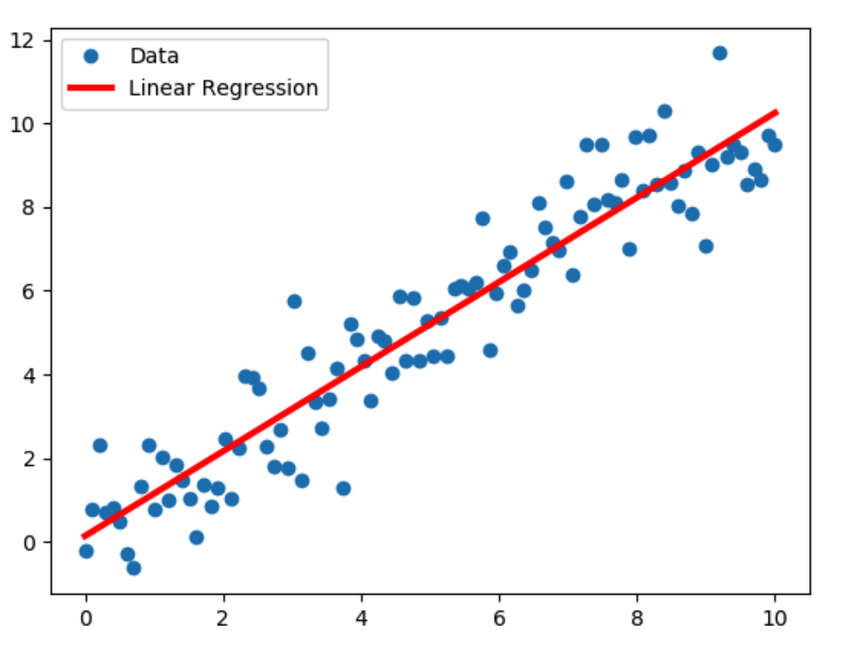
**Number of Admissions Over Time:**



**Detailed Information on Algorithms Used in Healthcare EDA**

**1. Linear Regression**

* Description: Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.
* Use Case in Healthcare: Predicting healthcare costs based on patient age and length of stay in the hospital.
* Advantages: Simple to implement, easy to interpret coefficients, works well with linearly separable data.
* Limitations: Assumes linearity, sensitive to outliers, and may not perform well with complex, non-linear relationships.



**2. Decision Tree Regression**

* Description: Decision Tree Regression splits the data into subsets based on the value of input features, and each split is chosen to minimize the prediction error.
* Use Case in Healthcare: Predicting healthcare costs by learning decision rules inferred from the features (e.g., age, length of stay).
* Advantages: Easy to understand and visualize, captures non-linear relationships, no need for feature scaling.
* Limitations: Prone to overfitting, sensitive to small variations in the data.

**3. Random Forest Regression**

* Description: Random Forest Regression is an ensemble method that uses multiple decision trees and aggregates their results to improve predictive accuracy and control overfitting.
* Use Case in Healthcare: More robust prediction of healthcare costs by combining multiple decision trees.
* Advantages: Reduces overfitting, handles large datasets with higher dimensionality, improves accuracy.
* Limitations: Complex and computationally intensive, less interpretable than individual decision trees.

**4. Polynomial Regression**

* Description: Polynomial Regression extends linear regression by introducing polynomial terms of the input features to model non-linear relationships.
* Use Case in Healthcare: Capturing the non-linear relationship between healthcare costs and features like age and length of stay.
* Advantages: Can model non-linear relationships, relatively simple extension of linear regression.
* Limitations: Prone to overfitting with high-degree polynomials, can become complex and computationally expensive.

**5. Ridge Regression**

* Description: Ridge Regression is a type of linear regression that includes a regularization term to penalize large coefficients, reducing model complexity and multicollinearity.
* Use Case in Healthcare: Providing robust predictions of healthcare costs while handling multicollinearity.
* Advantages: Prevents overfitting, handles multicollinearity, improves model generalization.
* Limitations: Requires tuning of the regularization parameter, less interpretable than standard linear regression.

**6. Lasso Regression**

* Description: Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds a penalty equal to the absolute value of the magnitude of coefficients, encouraging sparsity in the model.
* Use Case in Healthcare: Selecting significant features and predicting healthcare costs.
* Advantages: Performs feature selection, reduces overfitting, improves model interpretability.
* Limitations: Can be computationally intensive, might exclude important features if the penalty is too strong.

**Distribution of Data Using Gender in Power BI**

**Graph Description:**

A bar chart is used to display the distribution of healthcare data by gender. The X-axis represents the gender categories (Male, Female), while the Y-axis indicates the count or proportion of patients in each category. The chart is color-coded for easy differentiation between the genders.

* **Gender Balance:**

The graph shows the total number of male and female patients within the dataset. If the bars are of equal height, it indicates a balanced representation of both genders. Discrepancies can highlight gender-based differences in healthcare access or disease prevalence.

* **Healthcare Utilization:**

A higher count for one gender may suggest differences in healthcare utilization. For example, if females have a higher count, it might indicate more frequent hospital visits or higher health issues prevalence among women in the dataset.

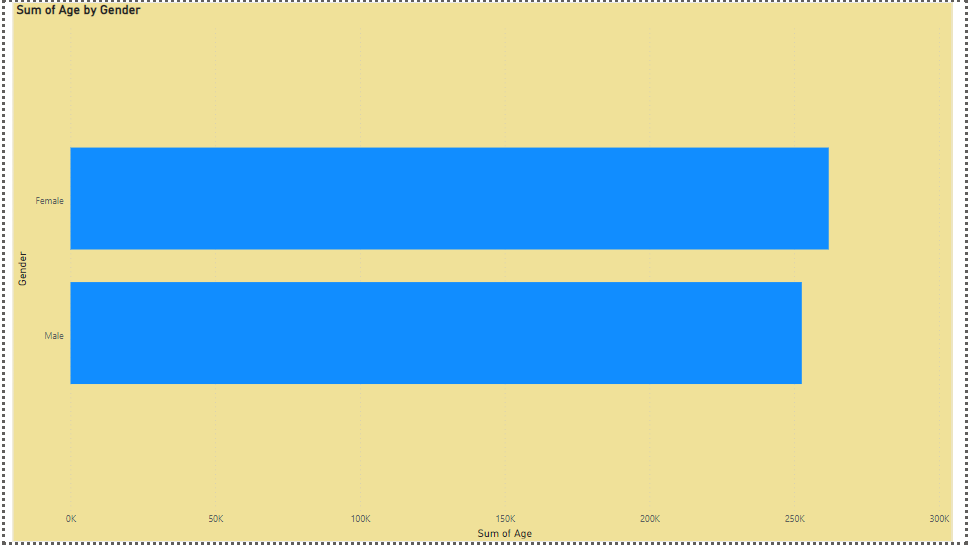


Figure 4.1Distribution of Data Using Gender in Power BI

**Bar Graph Description: Disease According to Age**

**Graph Description:**

The bar graph represents the distribution of various diseases across different age groups. The X-axis lists age groups (e.g., 0-20, 21-40, 41-60, 61+), and the Y-axis shows the count or proportion of patients diagnosed with specific diseases within each age group. Different colors or segments within each bar may indicate different diseases.

**Age-Specific Disease Prevalence:**

The graph highlights which diseases are more prevalent in specific age groups. For instance, chronic diseases such as hypertension and diabetes may show higher counts in older age groups, while infectious diseases might be more common in younger age groups.

**Healthcare Planning:**

Identifying age groups with higher disease prevalence helps healthcare providers plan and allocate resources effectively. For example, if a certain age group has a high prevalence of diabetes, more diabetes screening and management programs can be implemented for that demographic.

**Preventive Healthcare:**

By understanding which diseases affect certain age groups more, preventive measures and health awareness campaigns can be targeted appropriately. For instance, younger age groups may benefit from vaccination drives, while older age groups might require more lifestyle management programs.

**Policy Development:**

Policymakers can use this information to develop age-specific health policies. For example, special health insurance plans or wellness programs can be designed to cater to the needs of different age groups based on their prevalent diseases.

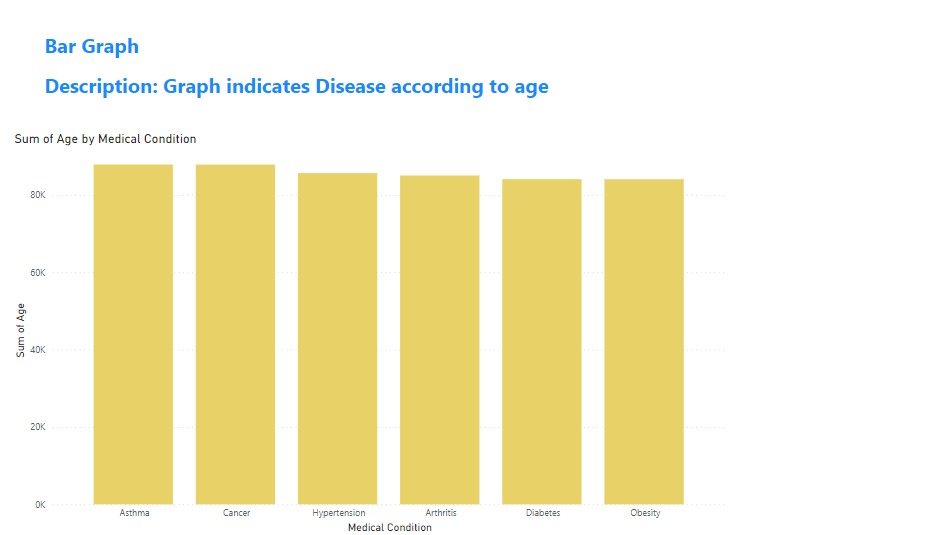


Figure 4.2Bar Graph Description: Disease According to Age

**Pie Graph Description: Count of Hospitals by Insurance Provider**

**Graph Description:** A pie chart displays the distribution of hospitals based on the insurance providers they are associated with. Each slice of the pie represents a different insurance provider, and the size of each slice corresponds to the count of hospitals accepting that insurance.

1. **Insurance Provider Market Share:**
   * The pie chart reveals the market share of different insurance providers within the hospital network. Larger slices indicate a higher number of hospitals accepting a particular insurance provider, reflecting their market dominance.
2. **Patient Accessibility:**
   * Understanding the distribution helps in assessing patient accessibility to healthcare services covered by their insurance. If one provider has a significantly larger share, patients with that insurance have more options for hospital care.

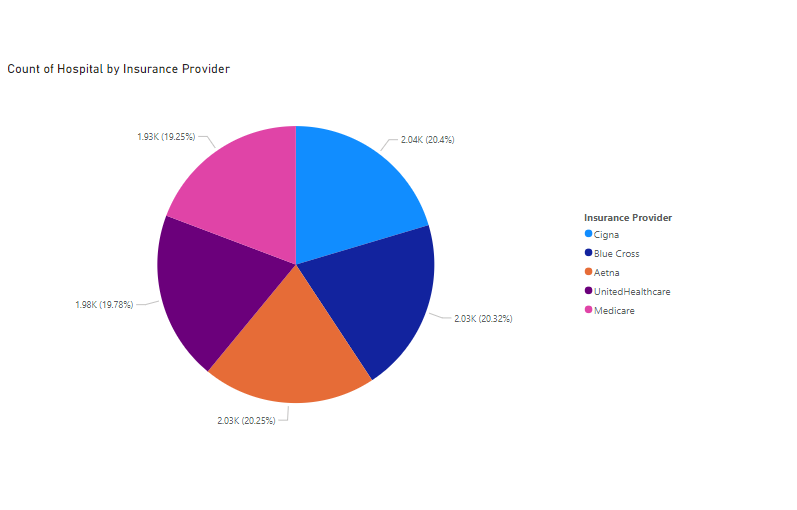


Figure 4.3Pie Graph Description: Count of Hospitals by Insurance Provider

**Line graph representing test results by medical conditions**

1. **Line Graph Representation**: A line graph typically plots data points over a continuous interval, often showing trends or changes over time. In the context of medical conditions, it might represent the frequency or count of test results associated with different conditions over a period of time.
2. **Key Points**:

**X-Axis**:

This would likely represent the medical conditions being tested for. Each condition would have its own data point or category along this axis.

**Y-Axis**:

This would represent the count or frequency of test results. It could be measured in absolute numbers or percentages, depending on the context.

**Lines**:

Each line on the graph would represent a specific medical condition. The slope and direction of the line indicate whether the frequency of test results for that condition is increasing, decreasing, or remaining stable over time.

**Data Points**:

These are the individual points plotted on the graph, representing the count of test results for each medical condition at specific time interval

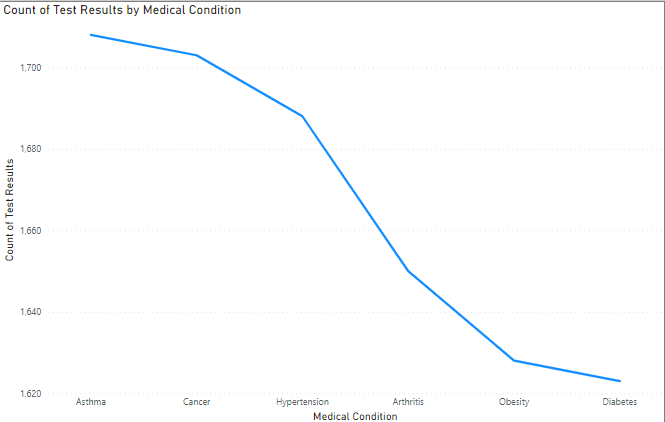


Figure 4.4Line graph representing test results by medical conditions

**Pie chart distribution using type of blood**

1. **Pie Chart Representation**: A pie chart is a circular statistical graphic that is divided into slices to illustrate numerical proportions. Each slice represents a proportionate part of the whole. In the context of blood types, it would represent the distribution of different blood types within a population or dataset.
2. **Key Points:**
   * Slices: Each slice of the pie chart represents a different blood type. Common blood types include A, B, AB, and O.
   * Proportions: The size of each slice corresponds to the proportion of individuals within the population or dataset that have that particular blood type. For example, if 40% of the population has blood type A, then the slice representing blood type A would occupy 40% of the pie chart.
   * Labels: Each slice is usually labelled with the name or abbreviation of the blood type it represents, along with the percentage or count it represents.
3. **Explanation of Graph:**
   * The pie chart visually represents the distribution of blood types within a population or dataset.
   * It provides a clear illustration of the relative prevalence of each blood type.
   * By comparing the sizes of the slices, viewers can quickly understand which blood types are more common and which are less common within the population.
   * The pie chart helps in understanding the demographic composition of blood types, which can be crucial for medical purposes such as blood transfusions or organ donations.
   * Overall, the pie chart offers a concise and easily interpretable summary of the distribution of blood types, making it useful for both medical professionals and the general public in understanding population health characteristics.

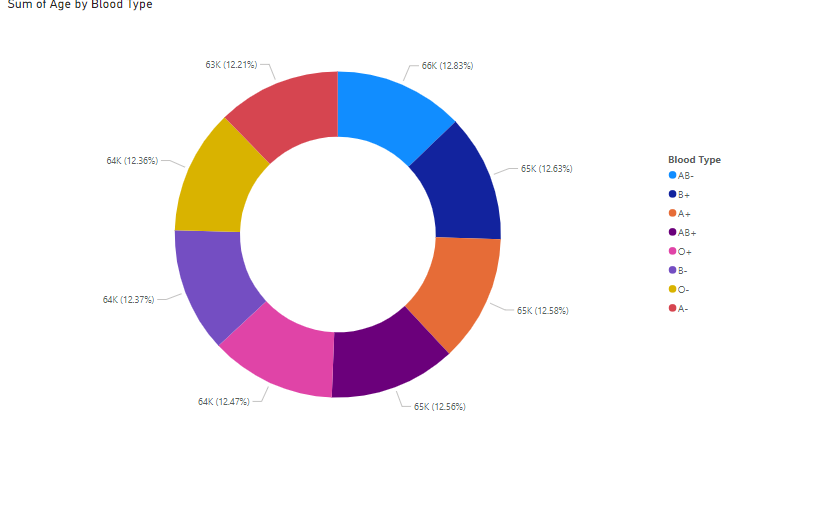


Figure 4.5Pie chart distribution using type of blood

**Pie graph representing the count of hospitals based on medical conditions**

1. **Pie Chart Representation**: Similar to the previous explanation, a pie chart is a circular statistical graphic divided into slices to illustrate numerical proportions. In this case, each slice represents the count or proportion of hospitals categorized by specific medical conditions.
2. **Key Points**:
   * **Slices**: Each slice of the pie chart represents a different medical condition. For instance, slices could include conditions like cardiovascular diseases, respiratory disorders, infectious diseases, etc.
   * **Proportions**: The size of each slice corresponds to the proportion of hospitals that primarily focus on treating patients with that specific medical condition. Larger slices indicate a higher number of hospitals specializing in that area, while smaller slices represent fewer hospitals.
   * **Labels**: Each slice is typically labeled with the name or abbreviation of the medical condition it represents, along with the percentage or count of hospitals specializing in that area.
   * **Total**: The entire pie chart represents the total number of hospitals being analyzed. The sum of the percentages of all the slices should equal 100%.
3. **Explanation of Graph**:
   * The pie chart visually represents how hospitals are distributed based on their specialization in treating specific medical conditions.
   * It offers a clear illustration of the relative prevalence of hospitals catering to different medical needs within a region or healthcare system.
   * By examining the sizes of the slices, viewers can quickly discern which medical conditions have a greater concentration of specialized hospitals and which may be relatively underserved.
   * The pie chart aids in understanding the healthcare infrastructure's allocation and distribution, highlighting areas of strength and potential gaps in medical services.
   * Overall, the pie chart serves as a useful tool for healthcare planners, policymakers, and the general public in assessing the distribution and accessibility of specialized medical care across different medical conditions within a given healthcare system or region.

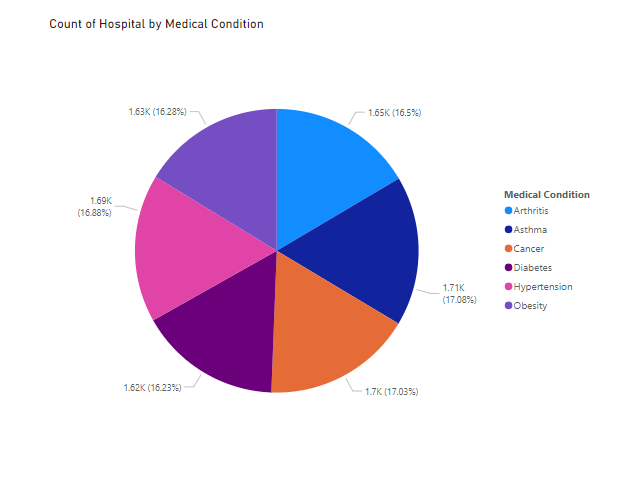
****

Figure 4.6Pie graph representing the count of hospitals based on medical conditions

|  |  |
| --- | --- |
| |  | | --- | | CHAPTER 5 | |

# CODE AND IMPLEMENTATION

Following is the code with explanation

## Loading and analyzing data

|  |
| --- |
| import pandas as pd  import seaborn as sns  import matplotlib.pyplot as plt  data = pd.read\_csv('healthcare\_dataset.csv')  print("First few rows of the dataset:")  print(data.head()) |

## Splitting the data in train and test set

1. **The Importance of Data Splitting:**

* Supervised machine learning is about creating models that precisely map the given inputs (independent variables, or predictors) to the given outputs (dependent variables, or responses).
* How you measure the precision of your model depends on the type of a problem you’re trying to solve. In regression analysis, you typically use the coefficient of determination, root mean-square error, mean absolute error, or similar quantities. For classification problems, you often apply accuracy, precision, recall, F1 score, and related indicators.
* The acceptable numeric values that measure precision vary from field to field.
* What’s most important to understand is that you usually need unbiased evaluation to properly use these measures, assess the predictive performance of your model, and validate the model.
* This means that you can’t evaluate the predictive performance of a model with the same data you used for training. You need evaluate the model with fresh data that hasn’t been seen by the model before. You can accomplish that by splitting your dataset before you use it.

## Training, Validation, and Test Sets

Splitting your dataset is essential for an unbiased evaluation of prediction performance. In most cases, it’s enough to split your dataset randomly into three subsets:

**The training set** is applied to train, or fit, your model. For example, you use the training set to find the optimal weights, or coefficients, for linear regression, logistic regression, or neural networks.

**The validation set** is used for unbiased model evaluation during hyperparameter tuning. For example, when you want to find the optimal number of neurons in a neural network or the best kernel for a support vector machine, you experiment with different values. For each considered setting of hyperparameters, you fit the model with the training set and assess its performance with the validation set.

**The test set** is needed for an unbiased evaluation of the final model. You shouldn’t use it for fitting or validation.

|  |
| --- |
| #Creating function to Split the data import pandas as pd  import numpy as np  import matplotlib.pyplot as plt  import seaborn as sns  from sklearn.model\_selection import train\_test\_split  from sklearn.linear\_model import LinearRegression  from sklearn.tree import DecisionTreeRegressor  from sklearn.ensemble import RandomForestRegressor  from sklearn.preprocessing import PolynomialFeatures  from sklearn.pipeline import make\_pipeline  from sklearn.metrics import mean\_squared\_error, r2\_score  def load\_and\_preprocess\_data(file\_path):  df = pd.read\_csv(file\_path)    # Convert date columns to datetime  df['Date of Admission'] = pd.to\_datetime(df['Date of Admission'], format='%Y-%m-%d', errors='coerce')  df['Discharge Date'] = pd.to\_datetime(df['Discharge Date'], format='%Y-%m-%d', errors='coerce')    # Calculate Length of Stay  df['Length of Stay'] = (df['Discharge Date'] - df['Date of Admission']).dt.days    # Handle missing or incorrect values if any  df.dropna(inplace=True)    return df  def perform\_eda(df):  # Summary Statistics  print("Summary Statistics:")  print(df.describe()) |

|  |
| --- |
| #calling function to split the data  def train\_and\_evaluate\_models(df):  # Selecting features and target variable for regression  features = df[['Age', 'Length of Stay']]  target = df['Billing Amount']    # Split the data into training and testing sets  X\_train, X\_test, y\_train, y\_test = train\_test\_split(features, target, test\_size=0.2, random\_state=42)    # Create regression models  models = {  'Linear Regression': LinearRegression(),  'Decision Tree Regression': DecisionTreeRegressor(random\_state=42),  'Random Forest Regression': RandomForestRegressor(n\_estimators=100, random\_state=42),  'Polynomial Regression (degree=2)': make\_pipeline(PolynomialFeatures(degree=2), LinearRegression()),  'Polynomial Regression (degree=3)': make\_pipeline(PolynomialFeatures(degree=3), LinearRegression())  }    # Train and evaluate each model  results = {}  for name, model in models.items():  model.fit(X\_train, y\_train)  y\_pred = model.predict(X\_test)  mse = mean\_squared\_error(y\_test, y\_pred)  r2 = r2\_score(y\_test, y\_pred)  results[name] = {'MSE': mse, 'R^2': r2}  print(f"{name} - Mean Squared Error: {mse}, R^2 Score: {r2}")    return models, results, X\_test, y\_test |

## Looking for Correlations :

The correlation matrix is a matrix that shows the correlation between variables. It gives the correlation between all the possible pairs of values in a matrix format.

We can use a correlation matrix to summarize a large data set and to identify patterns and make a decision according to it. We can also see which variable is more correlated to which variable, and we can visualize our results.

A correlation matrix involves a rows and columns table that shows the variables. Every cell in a matrix contains the correlation coefficient. The correlation matrix is in conjunction with other types of statistical analysis.

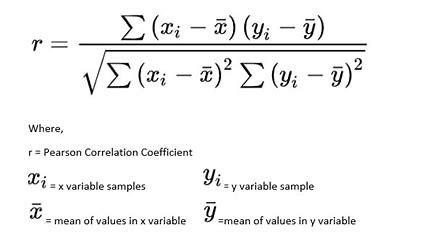
It’s very useful for regression techniques like simple linear regression, multiple linear regression and lasso regression models. In the regression technique, we have several independent variables, and based on that, we are predicting the dependent variable.

The correlation matrix gives you an idea about your data set.

For example, let’s say you want to predict the price of a car on the basis of fuel type, transmission type or age, etc. A correlation matrix would be very useful.

Using a correlation matrix, we can evaluate the relationship between two variables:

* If the relationship is 1, then the relationship is strong.
* If the relationship is 0, then it means the relationship is neutral.
* If the relationship is -1, then it means the relationship is negative or not strong.
* By using a correlation matrix, you can better understand your data set, analyse it and visualize the result.
* Most data scientists consider this the main step before building any machine learning model because if you know which variables are correlated which, you can gain a better understanding about what’s most important for your model.
* The correlation matrix is a statistical technique that gives you the values between -1 to 1 which you can determine the relationship between variables

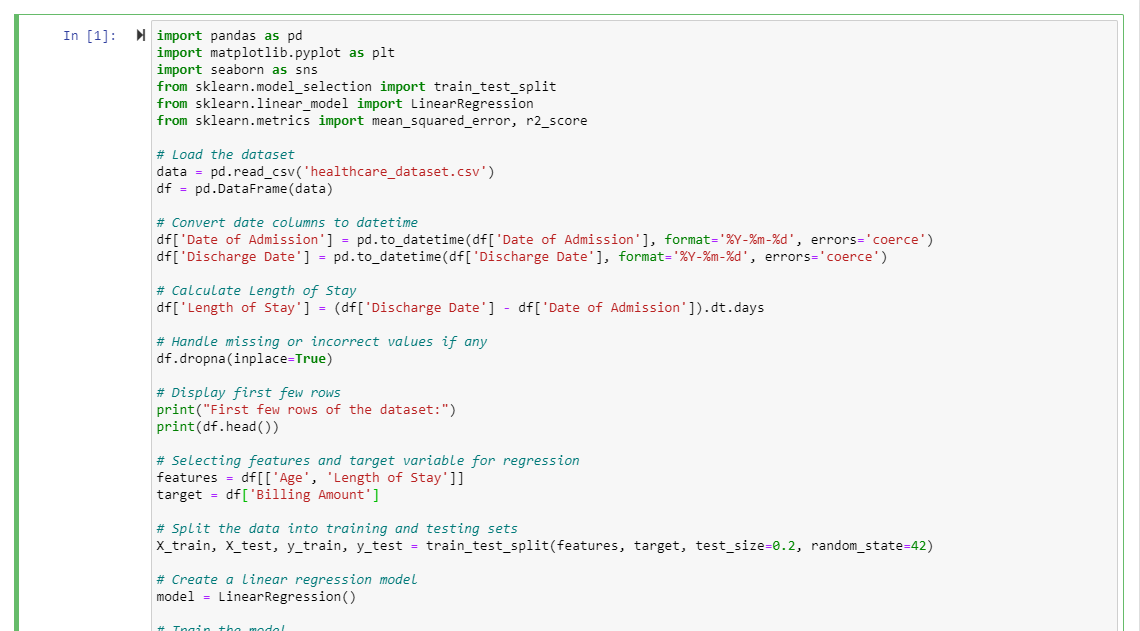


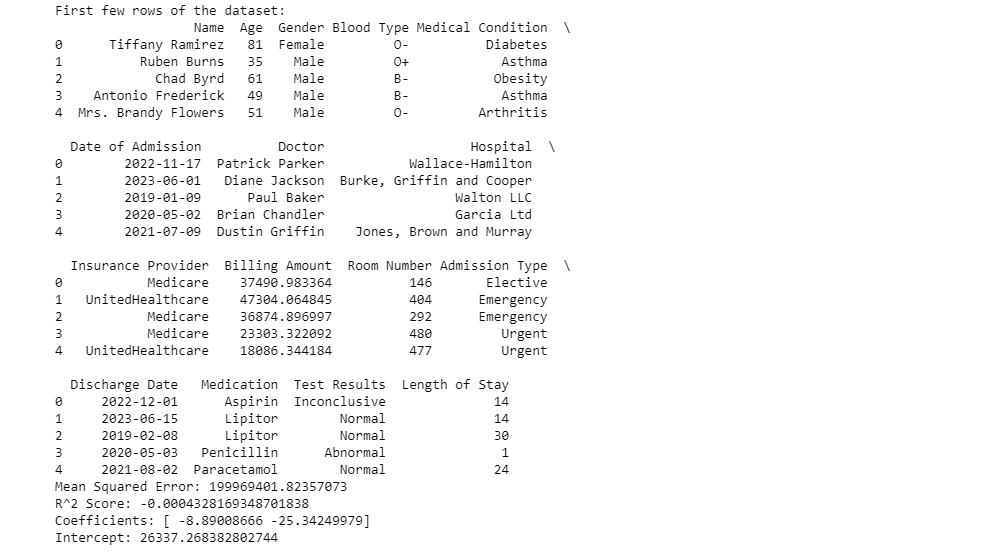
Drawing a heatmap for all the variables or features of the data go get an insight about the relation between them as we can observe same feature on X and Y axis gives co relation coefficients as 1 because the are totally same .

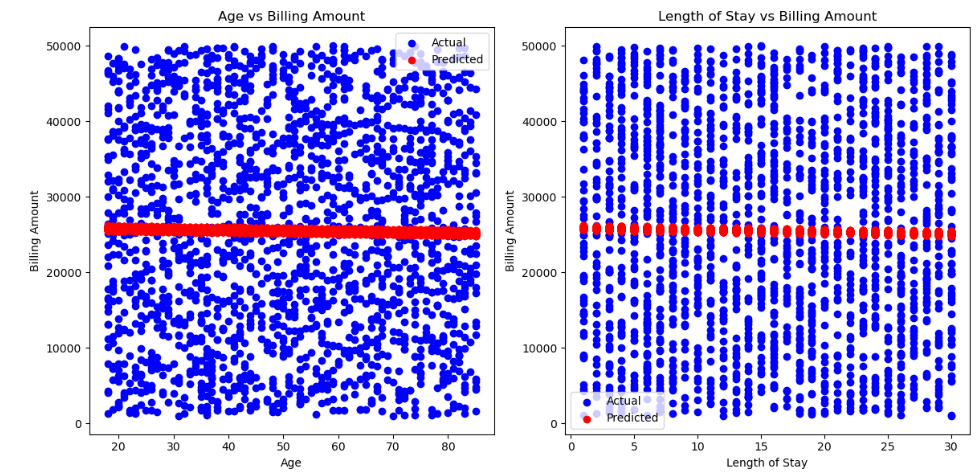
## Analysis using scatter plot between different attributes :

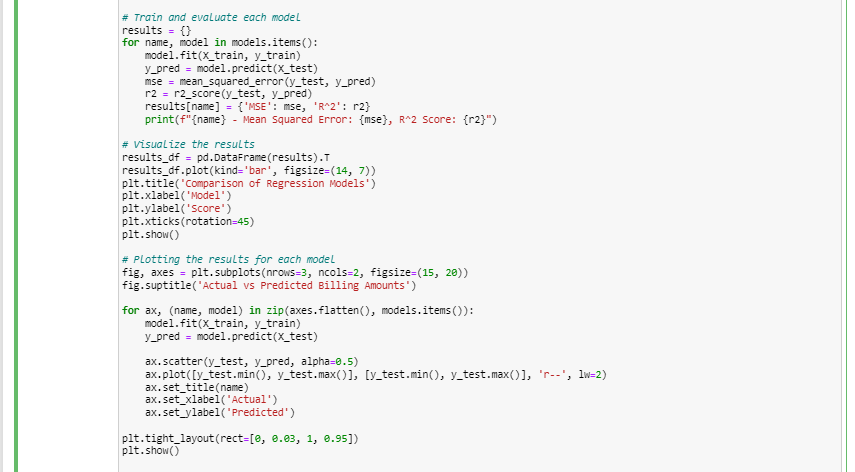
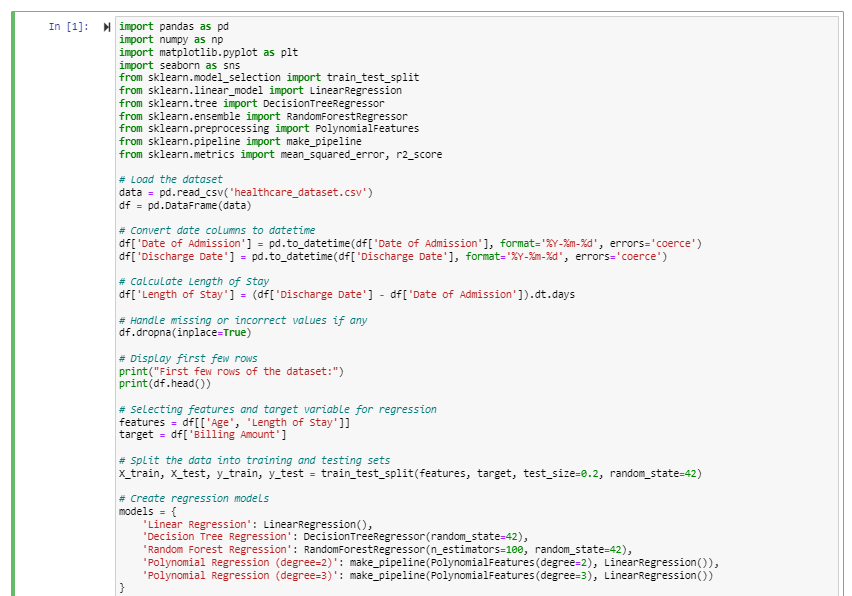
Scatter plots are used to observe relationship between variables and uses dots to represent the relationship between them. The scatter() method in the matplotlib library is used to draw a scatter plot. Scatter plots are widely used to represent relation among variables and how change in one affects the other. to get a general idea about whether or not two variables are related, is to plot them on a “scatter plot”. While there are many measures of association for variables which are measured at the ordinal or higher level of measurement, correlation is the most commonly used approach.

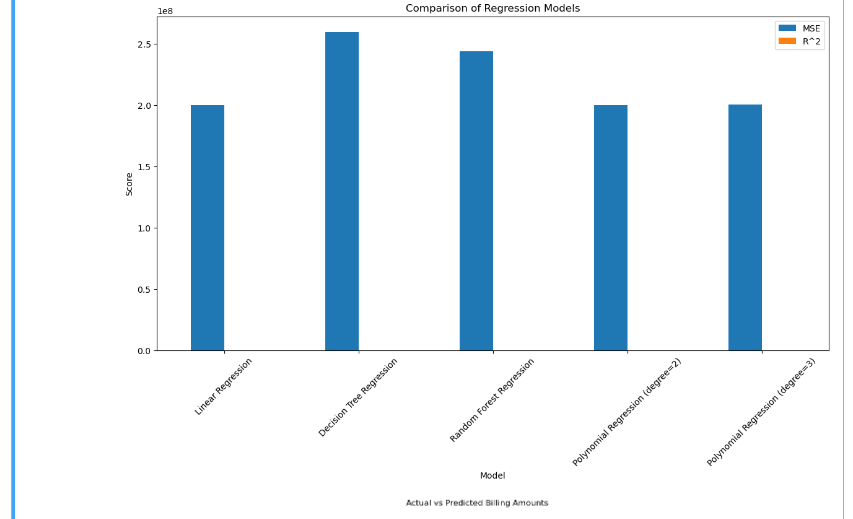
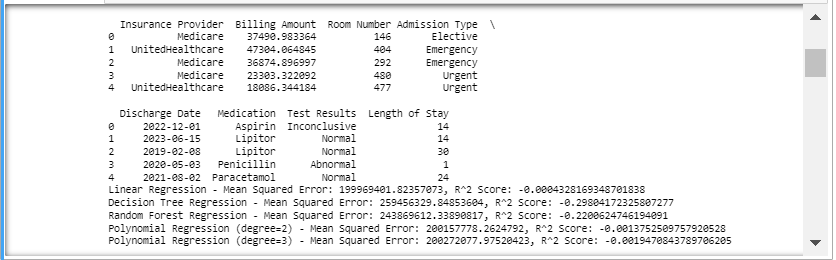
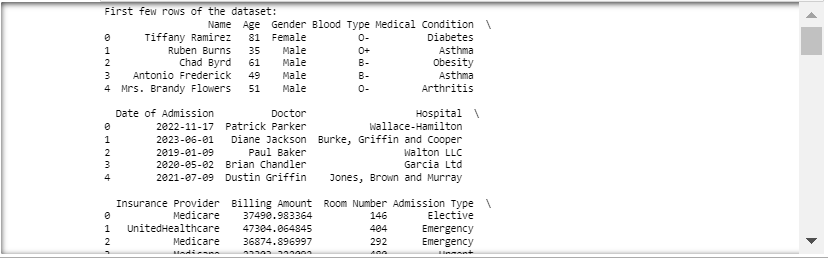
Below is the implementation

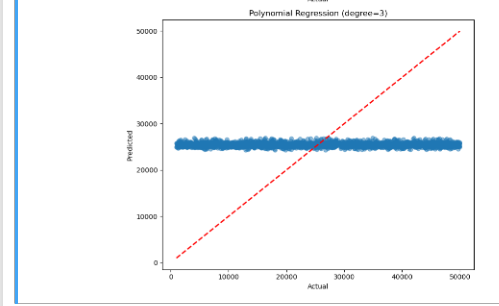
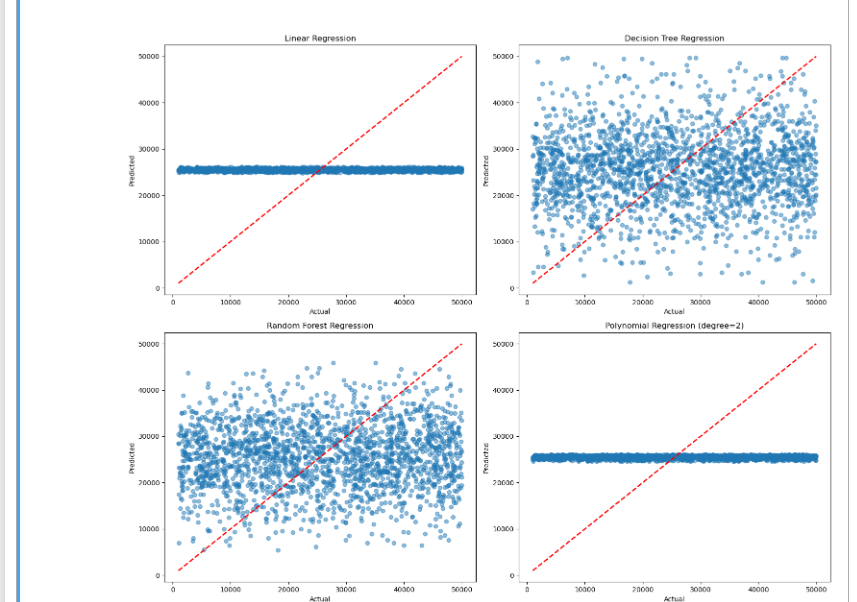
  
 





**Code Snip**

Output: 



|  |  |
| --- | --- |
| |  | | --- | | CHAPTER 6 | |

# CONCLUSION

**1. Introduction:**

* Brief overview of the importance of geolocational data in healthcare research.
* Explanation of exploratory analysis and its significance in uncovering insights from geospatial healthcare data.

**2. Geolocational Data in Healthcare:**

* Definition of geolocational data and its relevance in healthcare.
* Overview of various sources of geolocational data in healthcare, such as electronic health records (EHRs), patient-generated data, and environmental factors.
* Discussion on the potential applications of geolocational data in healthcare, including disease surveillance, public health interventions, and personalized medicine.

**3. Exploratory Analysis Techniques:**

* Explanation of exploratory data analysis (EDA) techniques commonly used in geospatial healthcare research.
* Overview of visualization methods, such as scatterplots, choropleth maps, and heatmaps, for exploring spatial patterns and relationships.
* Description of statistical analyses, including spatial autocorrelation, hotspot detection, and cluster analysis, for identifying spatial trends and clusters in healthcare data.

**4. Case Studies and Applications:**

* Review of published studies applying exploratory analysis to geolocational healthcare data.
* Case studies demonstrating the use of EDA techniques to analyze spatial patterns of diseases, healthcare access disparities, and environmental health risks.
* Discussion on the insights gained from these studies and their implications for healthcare policy and practice.

**5.Conclusion:**

The aim of the project is fulfilled successfully as we are able to predict the Diseases in healthcare domain and monitor the data using a machine learning model. Different steps performed in this problem solving are data extraction, data cleaning, data modelling, identifying missing values using the metrics calculation, filling the missing values, splitting the train and test data, choosing a machine learning model and finally predicting values of prices and evaluating the models using error functions and cross validation. Hence the project was implemented successfully.

**6. Challenges and Limitations:**

* Geolocational healthcare data.
* Discussion on issues such as data quality, privacy concerns, and interpretation of spatial patterns.
* Exploration of potential biases and confounding factors that may influence the results of geospatial analysis in healthcare research.

**7. Future Directions and Research Opportunities:**

* Proposal of future research directions and opportunities for advancing exploratory analysis of geolocational healthcare data.
* Suggestions for integrating advanced data analytics techniques, such as machine learning and spatial modelling, into geospatial healthcare research.
* Discussion on the potential impact of emerging technologies, such as wearable sensors and remote sensing, on the analysis of geolocational healthcare data.

BIBLIOGRAPHY

**BIBLIOGRAPHY**

Bibliography on Exploratory Data Analysis (EDA) in Healthcare

* Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

This seminal book by John Tukey introduced the concept of EDA, emphasizing the importance of exploring data to uncover underlying patterns before applying formal statistical methods. Tukey's work laid the foundation for modern EDA practices.

* Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). Understanding Robust and Exploratory Data Analysis. Wiley.

This edited volume expands on Tukey's original ideas, providing a comprehensive overview of robust and exploratory data analysis techniques, including practical applications in various fields.

Wilkinson, L. (2005). The Grammar of Graphics. Springer.

* Wilkinson's book is pivotal in the development of data visualization, an essential component of EDA. It provides a systematic approach to creating graphics that reveal data patterns, which is crucial for effective EDA in healthcare and other fields.

|  |  |
| --- | --- |
| |  | | --- | | REFERENCES | |

**REFERENCES**

* https://scikit-learn.org/
* Cross Validation in Machine Learning - GeeksforGeeks
* https://www.javatpoint.com/
* https://youtube.com
* https://www.kaggle.com/
* https://www.google.com
* Swayam - NPTEL

|  |  |
| --- | --- |
| |  | | --- | | ANNEXURE | |

# ANNEXURE A – QUESTIONNAIRE

Q1 . Need for predicting the diseases in healthcare?

Q2. Who is going to be benefitted by the output of the project?

Q3. Is data from reliable source?

Q4. Is it okay to assume values in place of missing values?

Q5. Is the model satisfying the needed results?

# ANNEXURE B – SCOPE FOR THE FUTURE

The Output can be very much useful and insightful for the stakeholders whereas as time goes on more features will be added to the database and the database will grow enormously in future so we need to devise more particular features and techniques to tune our algorithms to fit the data properly also more data cleaning and analysis will be needed in future to keep the predictions accurate and up-to-date.