EDA on Medical Appointments Data

Agenda

- Overview of the medical appointment scenerio
- About Dataset
- What is Exploratory Data Analysis?
- Why we need EDA?
- What is the approach of EDA?

Overview :

No-shows, or patients who miss their scheduled appointments are common and costly to healthcare institutions. A **US study** found that up to 30% of patients miss their appointments, and $150 billion is lost every year because of them.

Identifying potential no-shows can help healthcare institutions pursue targeted interventions (e.g. reminder phone calls, double-book an appointment slot) to reduce no-shows and financial loss.

‘Exploratory data analysis’ is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there.

Dataset:

The Kaggle dataset comprised 110k appointments records from public healthcare institutions in a Brazilian city. The appointments occurred across a 6-week period in 2016.

The **Exploratory Data Analysis** (EDA) is a set of approaches which includes univariate, bivariate and multivariate visualization techniques, dimensionality reduction, cluster analysis.

- The *main goal* of EDA is to get a **full understanding** of the data and draw attention to its most important features in order to prepare it for applying more advanced analysis techniques and feeding into **machine learning** algorithms. Besides, it helps to generate hypotheses about data, detect its anomalies and reveal the structure.

What EDA techniques are used :

The *graphical techniques* are the most natural for the human mind, therefore, plotting shouldn't be underestimated. These techniques usually include depicting the data using *box and whisker plots*, *histograms*, *lag plots*, *standard deviation plots*, *Pareto charts*, *scatter plots*, *bar and pie charts*, *violin plots*, *correlation matrices*, and more.

EDA approach:

1. Data Collection

2.Data Cleaning

3.Data Preprocessing

4. Data Visualisation

Data Collecting

Data collection is the process of gathering information in an established systematic way that enables one to test hypothesis and evaluate outcomes easily.

Data Cleaning

Data cleaning is the process of ensuring that your data is correct and useable by identifying any errors in the data, or missing data by correcting or deleting them.

Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. It includes normalisation and standardisation, transformation, feature extraction and selection, etc. The product of data preprocessing is the final training dataset.

Data Visualisation

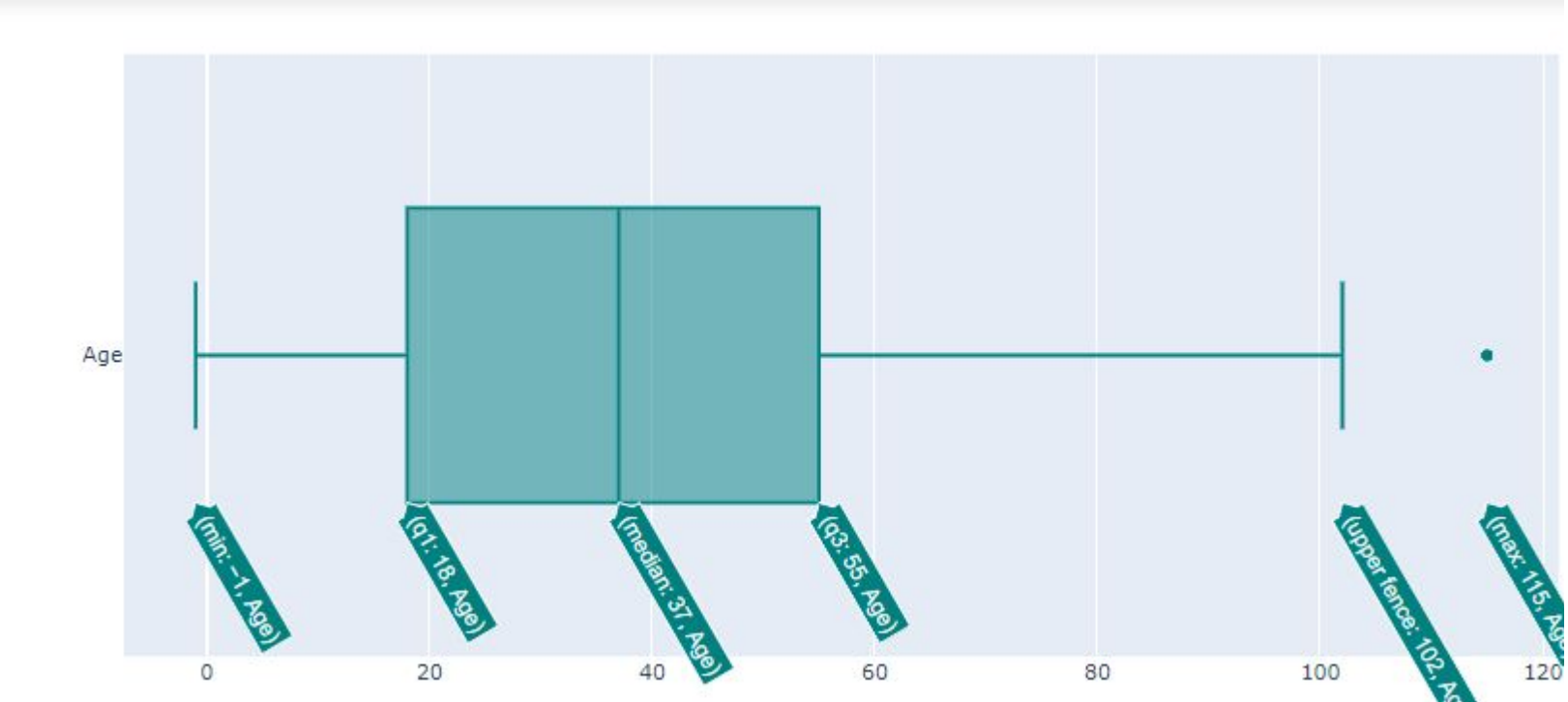
Data visualisation is the graphical representation of information and data. It uses statistical graphics, plots, information graphics and other tools to communicate information clearly and efficiently.

Following are the common used visualisation :-

- Scatter Plot : A scatter plot is a set of points plotted on a horizontal and vertical axes.
- Box Plot :
- Histogram
- Violin Plot
- Pair Plot
- Heat Map

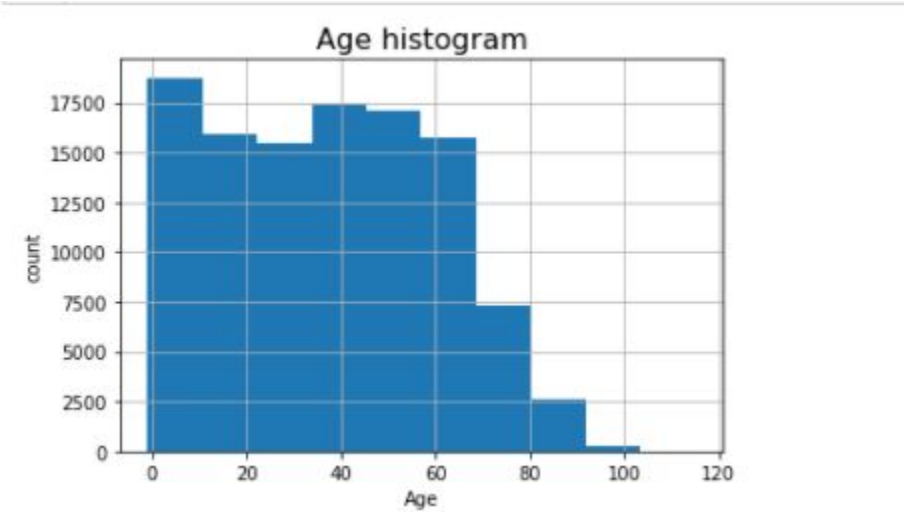
Box Plot

Box plot is a simple way of representing statistical data on a plot in which a rectangle is drawn to represent the second and third quartiles, usually with a vertical line inside to indicate the median value. The lower and upper quartiles are shown as horizontal lines either side of the rectangle.



Histogram

A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable.

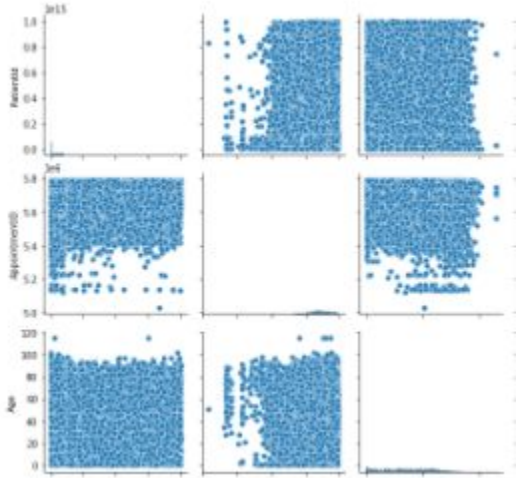


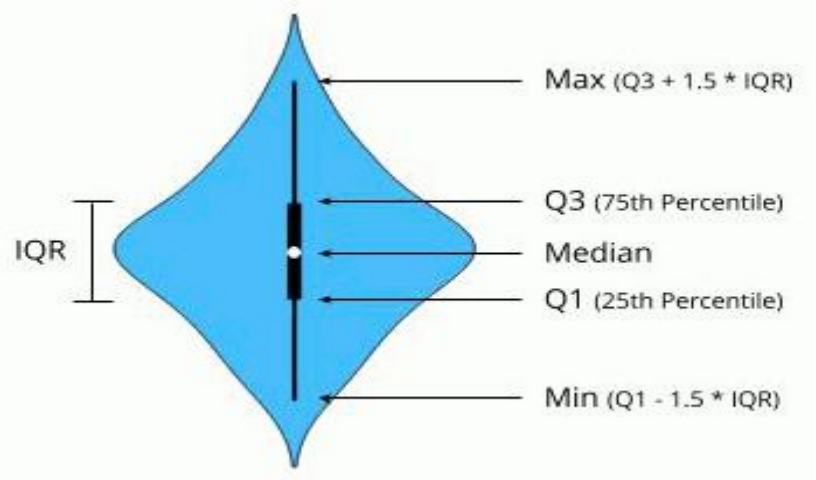
Violin Plot

A violin plot plays a similar role as a box and whisker plot. It shows the distribution of quantitative data across several levels of one (or more) categorical variables such that those distributions can be compared.

Pair Plot

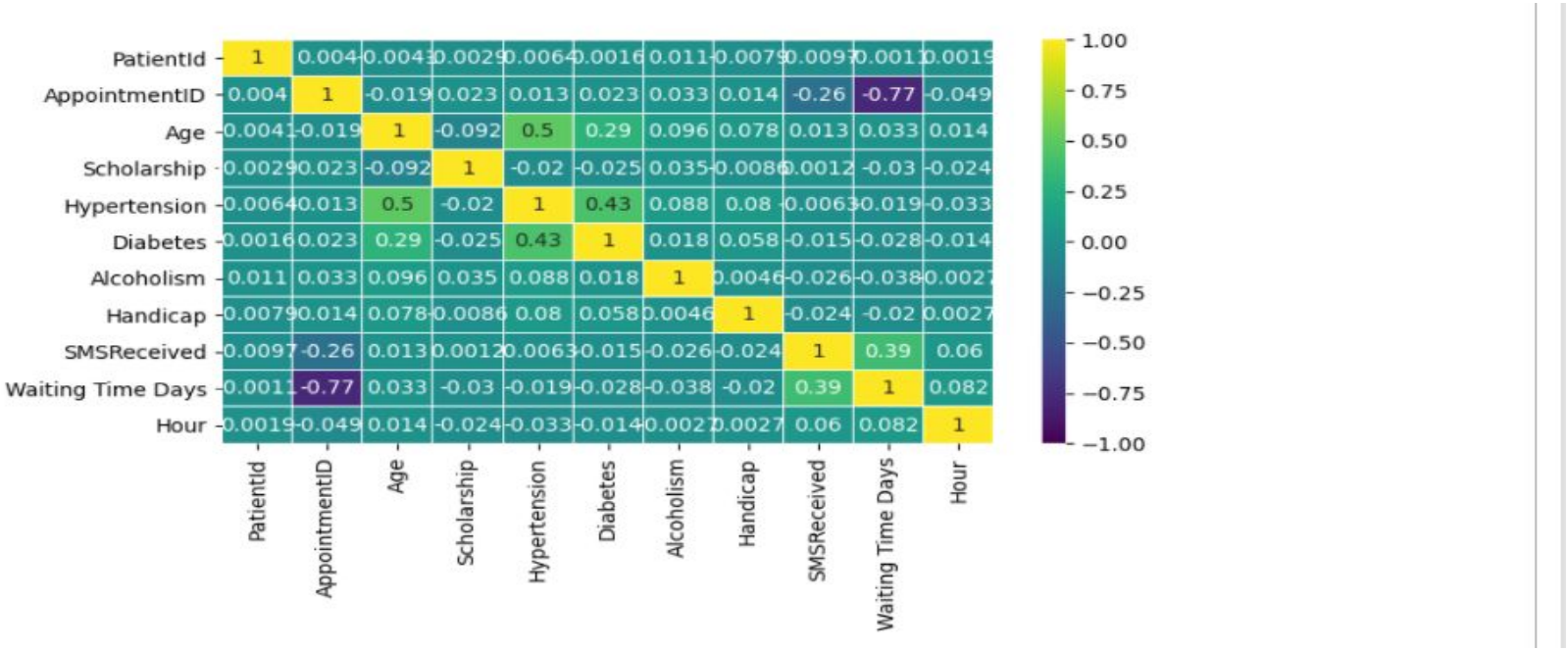
Pair plot in seaborn only plots numerical columns although later we will use the categorical variables for coloring.





Heat Map

Heat map is a representation of data in the form of a map or diagram in which data values are represented as colours.



Thank you