Meet Fatima

a keen data analyst, embarked on a journey in Exploratory Data Analysis (EDA) with a complex dataset aimed at guiding her company's strategies.

Understanding EDA as a crucial element of data science grounded in STEM, she utilized statistical techniques, computing technology, and an engineering approach to systematically solve problems, blending scientific theory with practical application to achieve the diverse objectives of EDA.

As she delved deeper, Fatima understood the multifaceted goals of EDA.

Each objective served as a guiding light in her exploration:

1. **Spotting Anomalies:**

Her initial task involved identifying outliers or unusual patterns in the data, which often revealed key insights or areas needing further examination.

2. **Hypothesis Testing in EDA**

allowed Fatima to test her assumptions about the data, a vital step for confirming or refuting her theories and understanding the dataset's characteristics.

3. Investigating Data:

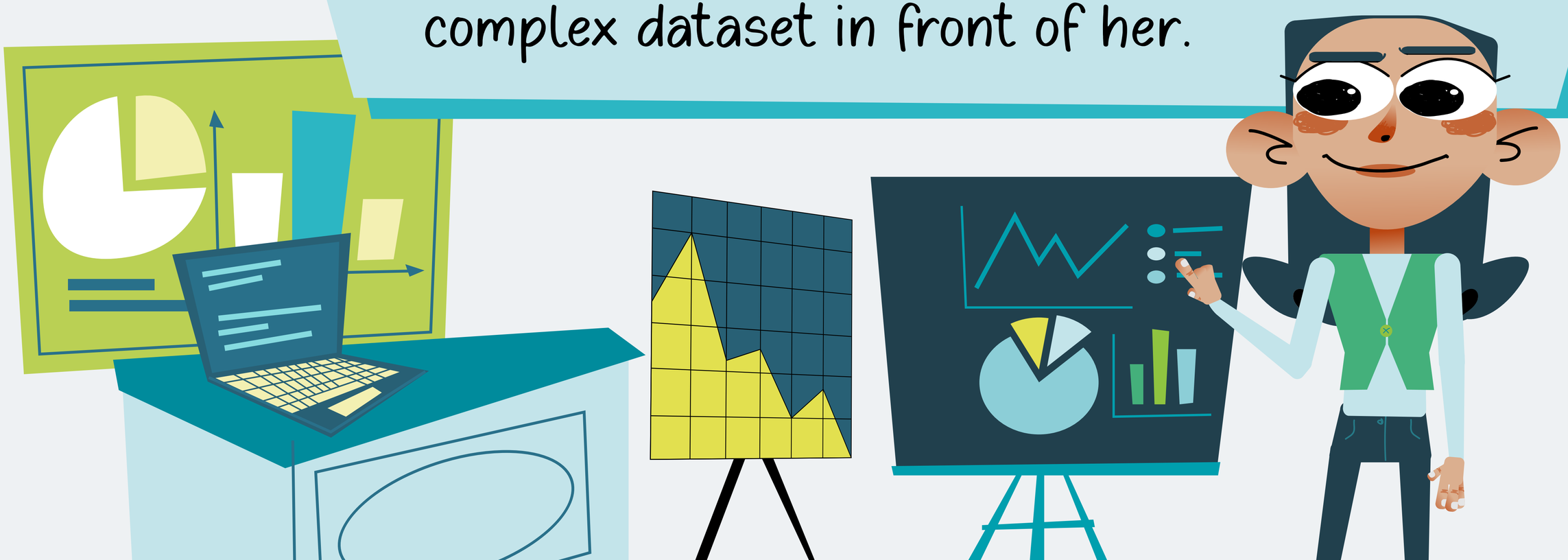
Fatima dedicated much time to deep data investigation, examining variables and their interrelations, akin to a detective piecing together clues to form a broader picture.

4. Discovering Patterns:

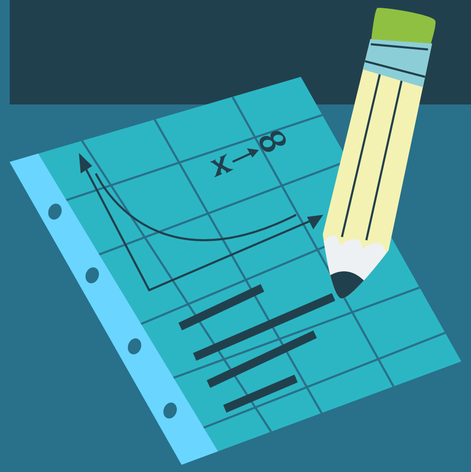
Fatima found excitement in uncovering patterns like trends, correlations, or groupings, which offered valuable insights and often directed further analysis.

For Fatima, embarking on her data analysis journey, descriptive statistics

were the first step in making sense of the complex dataset in front of her.

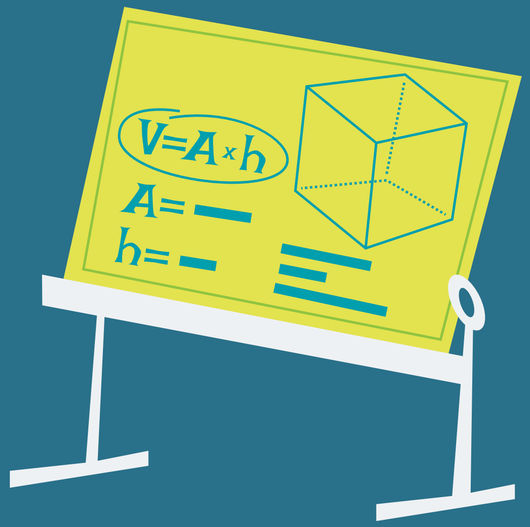


By applying these statistics, she gained a preliminary overview of her data:



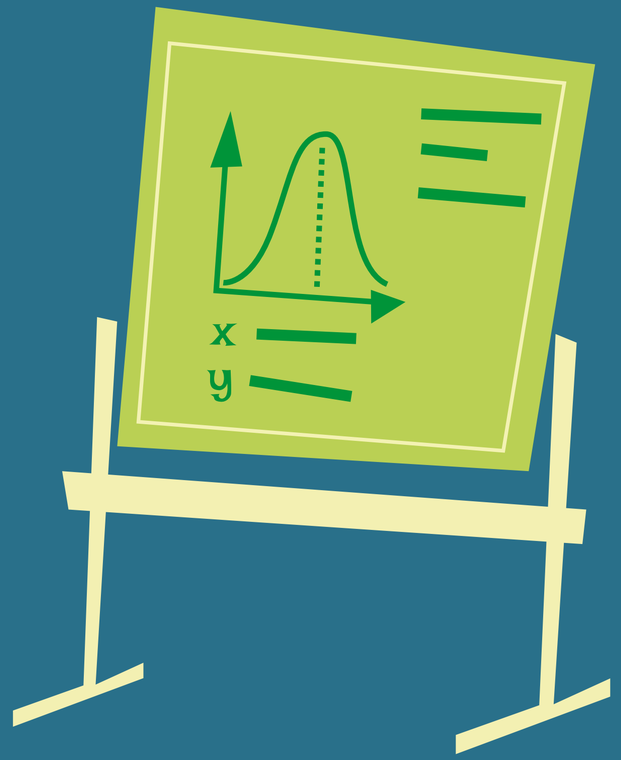
She calculated the mean and median to find the central tendency, which gave her a quick understanding of the average values in her dataset.

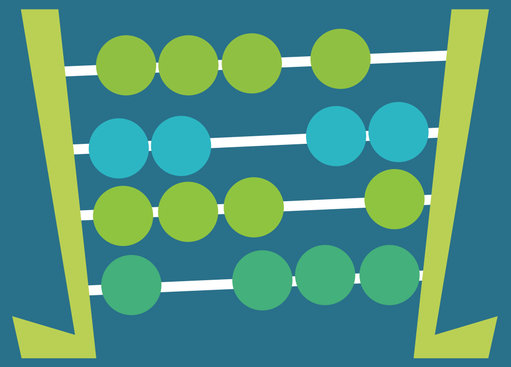
The range and standard deviation revealed how spread out her data points were, indicating the level of consistency in the data.



Examining the skewness and kurtosis, Fatima could infer if her data had any bias or were unusually peaked.

By analyzing quartiles and percentiles, she could identify outliers and understand the distribution of data across various thresholds.



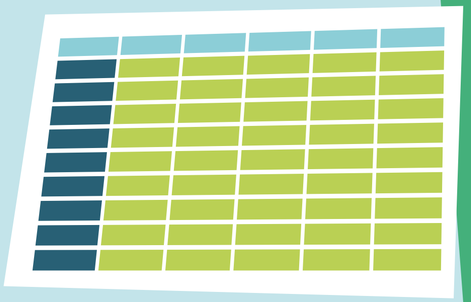


The frequency distributions helped her visualize the data, making it easier to spot patterns and anomalies.

Fatima conducted data diagnosis, akin to a doctor's examination, to identify any issues in the dataset that might impact the analysis's accuracy or reliability.

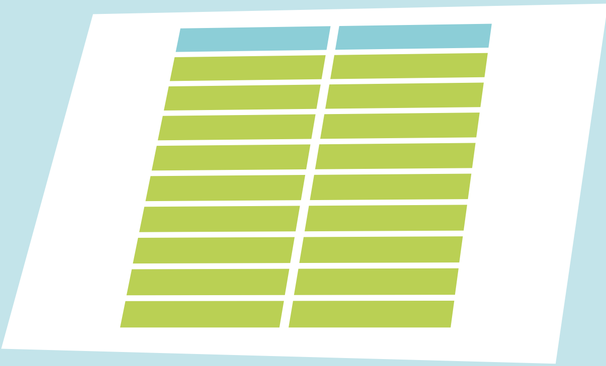
What are the Basic Steps of Diagnosis?

Fatima began by identifying each column's data type in her dataset, helping her understand whether the data is numerical, categorical, textual, or date/time.



She next cataloged the dataset's dimensions by counting rows and columns, providing a clear understanding of its size and scope for planning her analysis.

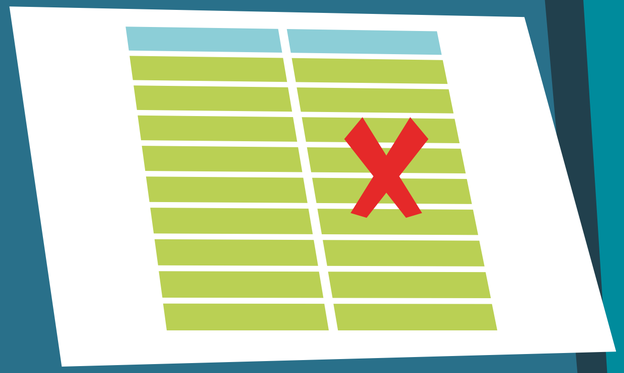
Column Data Type Specification: Next, Fatima specifies the exact data types of two specific columns, such as integer, floating-point number, string, or boolean.



Fatima checks for null or missing values in each column, a critical step for assessing data completeness and integrity.

Statistical Properties Evaluation: Finally, Fatima evaluates key statistical measures like mean, median, and variance for each column to understand data distribution and central tendencies, setting the stage for deeper analysis.

Fatima progresses to Data Cleaning after data diagnosis, aiming to enhance her dataset's quality.

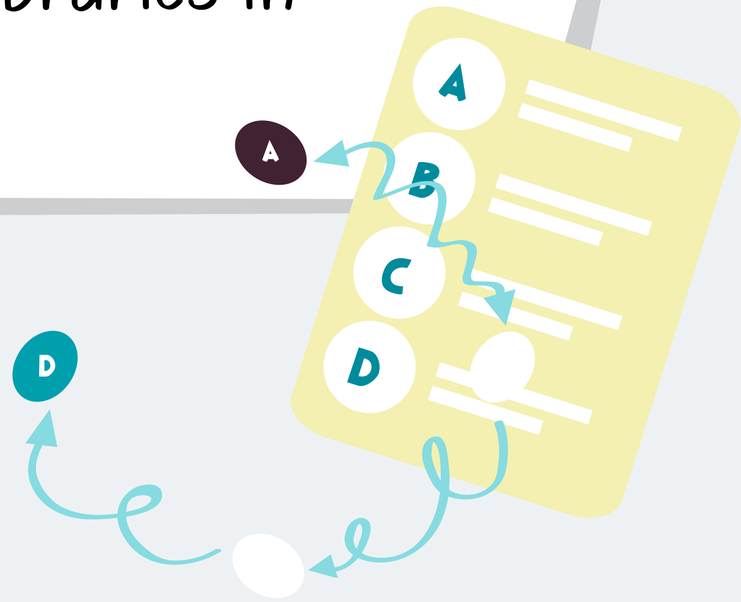


She starts by handling null values, deciding whether to remove or impute them. She then assesses each column's relevance, removing non-contributory ones.

Next, she identifies and eliminates duplicate records.

If Fatima faces challenges with the data type structure in her dataset, distinguishing between numerical and categorical types, she must correct these classifications.

Numerical data is essential for computational processes and optimizing memory in dataframes, crucial for libraries in statistical calculations and machine learning.



However, its value range can impact model performance, so normalization or scaling is employed to standardize ranges, balancing feature influence in analysis and model building. Normalization adjusts values within a [0, 1] range, benefiting scale-sensitive algorithms like K-Nearest Neighbors.

$$X_{\text{Normalized}} = \frac{X_{\text{intial}}}{\max(X)}$$

Standardization, on the other hand, to a mean of zero and standard deviation of one, ideal for methods like Support Vector Machines that assume normal distribution.

$$X_{Standardized} = \frac{X_{intial} - \max(X)}{std(X)}$$

$$X_{Normalized} = \frac{X_{intial}}{max\ (X)}$$

The choice between them depends on the specific algorithm and data characteristics, with normalization sensitive to outliers, whereas standardization is less so.

$$X_{Standardized} = \frac{X_{intial} - max\ (X)}{std\ (X)}$$

Fatima's transformed dataset, now a narrative beyond numbers and categories, reflects her skillful EDA journey. This process highlights the power of a methodical, insightful approach in uncovering hidden data stories.