Data Cleaning and Exploratory Data Analysis Report

I spent some time getting to know the BMI dataset, and here's a rundown of what I did, along with my key findings.

Data Cleaning

1. **Loading & Initial Inspection:**
    - I loaded the dataset and took a quick look at the first few rows and the overall structure. This helped me understand what kinds of data (numerical and categorical) I was dealing with.
2. **Handling Missing Values:**
    - I noticed that some cells were empty. For the numeric fields, I filled in the blanks with the median value to avoid distorting the data with extreme values. For text-based columns, I used the mode (the most common entry) to replace missing values. This step ensured that no data point was left hanging.
3. **Removing Duplicates:**
    - I checked for any duplicate records and removed them. Duplicates can really skew your analysis, so I made sure each record in the dataset was unique.
4. **Dealing with Outliers:**
    - To make sure that extreme values didn't throw off my analysis, I used the Interquartile Range (IQR) method to identify and remove outliers from the numerical columns. This process helped in cleaning up the data distribution.
5. **Standardizing Categorical Data:**
    - Finally, I standardized the text data by converting all entries to lowercase and trimming any extra spaces. This step helped ensure that entries like "Male" and "male" would be treated the same way in the analysis.

Exploratory Data Analysis (EDA)

1. **Univariate Analysis:**
    - **Summary Statistics:** I calculated key statistics (mean, median, variance, etc.) for each numerical variable. This gave me a clear picture of the central tendencies and spread of the data.

- **Visualizations:** I plotted histograms and box plots for each numerical column. These visuals were useful to quickly see the distribution and spot any remaining anomalies.

2. **Bivariate Analysis:**
    - I explored how pairs of variables related to each other:
        - **Correlation Matrix:** I computed the correlations between numerical variables and visualized them with a heatmap, which made it easy to spot strong relationships.
        - **Scatter and Box Plots:** I created scatter plots to check continuous relationships and used box plots to compare how numerical values, like BMI, vary across different categories (e.g., gender).
3. **Multivariate Analysis:**
    - **Pair Plots:** I generated pair plots to simultaneously look at relationships between multiple variables. This was especially helpful in understanding how variables interact in a broader context.
    - **Grouped Comparisons:** I grouped the data by key categorical variables (like gender) and calculated the average values for each group. For example, I examined how BMI differed between males and females.

Next Steps

- With the data now clean and insights from the EDA in hand, the dataset is well-prepared for further analysis. Whether you're planning to build predictive models or simply dive deeper into the data patterns, the groundwork is solid.
- One potential next step is to calculate the BMI (if the raw height and weight data are available) and then plot it for further insights.

Overall, the process involved methodically cleaning the data and using a mix of statistical summaries and visualizations to understand it. This approach not only improved data quality but also provided a comprehensive overview of underlying trends and patterns.

Thank you ,

Sushim Saini

23116096