**DATA Exploration**  
**Missing Values**

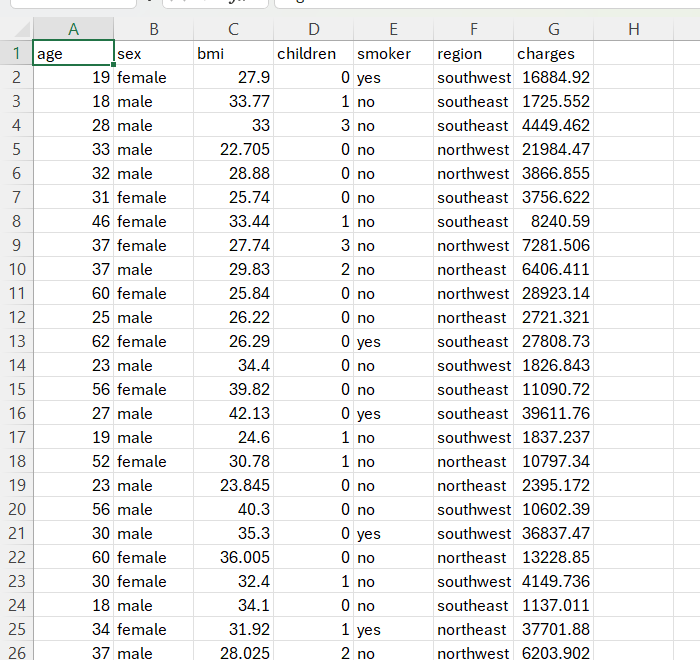
-> No missing Values were found on the data set

**Non-Numerical data**

-One hot encoded is used to converted categorical variable to data by creating new binary columns for each unique category value

-Good since, One-hot encoding does not assume any ordering among the categories, which is appropriate for nominal data.

Data before One hot encoding



Data after One hot encoding

**A screenshot of a computer

Description automatically generated**A black screen with white text

Description automatically generated

We created boxplots each column that are purely numerical to show the characteristics of the data individually. Plotting one hot encoded value in a box plot those not really make sense not unless you are relating it to another feature.

A screenshot of a graph

Description automatically generated

**1. Box Plot of age:**

* The age distribution is quite wide, ranging roughly from 20 to 65 years.
* The median age is around 39-40, indicating that half of the individuals are younger and half are older than this.
* There is a relatively balanced distribution with no visible outliers, suggesting that most ages fall within a common range without extreme cases.

**2. Box Plot of bmi:**

* The median BMI is approximately 30, which is categorized as overweight.
* There are several outliers beyond 40, which indicates some individuals have significantly higher BMIs (indicative of obesity).
* The interquartile range (IQR) for BMI shows a significant spread, meaning the BMI values vary widely among individuals.

**3. Box Plot of children:**

* The median number of children is 1, indicating that most individuals have at least one child.
* The majority of individuals have between 0 to 2 children, with a few having up to 5 children.
* The lower whisker extends to 0, which indicates that a significant proportion of individuals do not have any children.

**4. Box Plot of charges:**

* The distribution of charges is highly skewed, with several high-value outliers above 30,000.
* These outliers indicate a small number of individuals with much higher medical expenses, possibly due to specific health conditions.
* The median charges are relatively low compared to the maximum values, suggesting that while most individuals have moderate charges, there is a small group with much higher medical costs.
* The long tail and the clustering of outliers suggest that charges might not follow a normal distribution and are instead influenced by specific high-cost cases.

**General Insights:**

* **Outliers** are present in both bmi and charges, indicating that certain individuals have significantly different characteristics compared to the rest of the population, which could have implications in health and insurance costs.
* The data distributions, particularly for charges, are not symmetrical, suggesting that some of these variables might require transformations for statistical modeling to reduce skewness or improve model performance.
* **Age** and **children** have no outliers, meaning these features are relatively normally distributed and show less variability in extreme values compared to bmi and charges.

4o

**1. Box Plot of age:**

* There are **no outliers** visible in the age distribution.
* This suggests that all age values are within a common range without any extreme ages that are significantly different from the rest of the data.
* The distribution is balanced, and individuals’ ages fall within the whiskers without deviations.

**2. Box Plot of bmi:**

* There are **outliers** present beyond a BMI value of around 40.
* These outliers represent individuals with unusually high BMIs, which indicates obesity or extreme overweight. They might need further analysis to understand the underlying causes of these values.
* The presence of these outliers implies that there are some people whose BMIs are significantly different from the majority of the sample.

**3. Box Plot of children:**

* There are **no outliers** in the children data.
* This means that the number of children per individual falls within a typical range without extreme cases.
* The whiskers extend to the maximum number of children without indicating any unusually high values, which suggests that having up to 5 children is not considered uncommon for this dataset.

**4. Box Plot of charges:**

* There are **many outliers** for the charges feature, primarily above 30,000.
* These outliers indicate individuals with significantly higher medical expenses compared to the rest of the population. These could be due to chronic health issues, costly treatments, or other healthcare needs.
* The high number of outliers suggests a long-tailed distribution where a small group of individuals have disproportionately high healthcare costs, which might need special consideration during statistical analysis or modeling.

**General Insights Regarding Outliers:**

* The **bmi** and **charges** features have noticeable outliers. These outliers are important to understand, as they can have a significant impact on statistical models and may need to be treated differently (e.g., log transformation or exclusion) depending on the goal of the analysis.
* The **absence of outliers** in **age** and **children** suggests that these features are stable and don’t have extreme values, making them easier to work with in predictive modeling without needing special preprocessing steps for outlier handling.
* Outliers in **charges** might represent special cases (like specific health conditions) that could provide valuable insights if analyzed separately.