1. **Univariate analysis**

The most appropriate center is median because the distribution of charges of citizens of United States is skewed right. The most appropriate spread is IQR and there are no outliers. Whereas the most appropriate center is mean for the distribution of BMI of citizens of United States is normal. The most appropriate spread is standard deviation and there are no outliers. Coming to age, the most appropriate center is mean as it is distributed uniformly. The most appropriate spread is standard deviation, and it has no outliers.

Variables charges and age requires transformation because these two attributes doesn’t come from normal distribution. Box-cox transformation is used to transform the variables. After transformation distribution of charges looked like a bell curve whereas there is not much improvement in attribute age. Transformation is successful for ‘charges’.

**A graph of a number of objects

Description automatically generated with medium confidenceA graph of a number of charges

Description automatically generated with medium confidence**

**Table 1** **Numerical summaries for each variable**

|  |  |
| --- | --- |
| Variables | Summaries |
| Charges ($, ± sd) | 13270 ± 12110.01 |
| BMI (kg/m2, ± sd) | 30.66 ± 6.09 |
| Age (years, ± sd) | 39.21 ± 14.04 |

1. **Univariate analysis by groups**

The citizens who smokes are paying high insurance charges. People of southeast are relatively paying higher charges when compared to southwest and other regions. All the boxplots related to Charges are skewed right except for the ‘Charges by smoking people’ whose shape cannot be determined. And all these plots have ‘median’ as their center and ‘IQR’ as their spread. All the boxplots have outliers except ‘Charges by smoking people’.

BMI is respectively high in people of southeast relatively compared to southwest and other regions where other region recorded low BMI measures. From the box plots, we observe the BMI of smokers and non-smokers fall in same range. All the boxplots related to BMI are symmetric and all of them have ‘median’ as their center and ‘IQR’ as their spread. All of them has outliers.

From the box plots we observe that the Age of non-smokers is relatively higher than that of smokers with a slight margin. Whereas all three regions almost have same range of age. All the boxplots related to age are symmetric and has ‘median’ as their center and ‘IQR’ as their spread and all of them do not contain any outliers.

**Table 2.1 Numerical summaries by ‘Smoker’**

|  |  |  |
| --- | --- | --- |
| Variables | Smoking status | |
|  | Not a smoker | Smoker |
| Charges ($, ± sd) | 8434 ±5994 | 32050±11542 |
| BMI (kg/m2, ± sd) | 30.7±6.04 | 30.7±6.32 |
| Age (years, ± sd) | 39.4±14.1 | 38.5±13.9 |

**Table 2.2 Numerical summaries by ‘Region’**

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Region | | |
|  | Southeast | Southwest | Other |
| Charges ($, ± sd) | 14735 ±13971 | 12347 ± 11557 | 12911 ± 11167 |
| BMI kg/m2, ± sd) | 33.4 ± 6.48 | 30.6 ± 5.69 | 29.2 ± 5.55 |
| Age (years, ± sd) | 38.9 ± 14.2 | 39.5 ± 14.0 | 39.2 ± 14.0 |

1. **A screenshot of a graph

   Description automatically generatedBivariate Analysis**

|  |  |
| --- | --- |
| Smoker | |
| 0 | Not a smoker |
| 1 | Smoker |
| Region | |
| 1 | Southeast |
| 2 | Southwest |
| 3 | Other region |

**A screenshot of a graph

Description automatically generated**

As people get older, they tend to pay more for insurance. This is because older individuals often require more medical care, leading to higher insurance costs.

Age and charges have a positive linear relationship, however, the relationship isn't perfectly linear, as there are some outliers. Regarding charges and BMI, there's no apparent trend or direction in the scatter plot, indicating a weak or negligible relationship between the two variables. Similarly, when examining BMI and age, there's no clear pattern or direction, suggesting a weak correlation. Overall, while age appears to influence charges with a moderate strength, BMI doesn't seem to have a significant impact, and there are notable outliers in the data.

Scatterplot colored by different categorical variable suggests that smokers of southeast are paying higher charges along with smokers of other regions. Additionally, smokers experience a bigger increase in insurance charges as they age compared to non-smokers. Where people live(region) don’t seem to have a big impact on how much they pay for insurance. However, nonsmokers tend to pay low fares. Overall, smoker status is playing a key role over the regions in determining the insurance charges.

1. **Multivariate Analysis**

**A diagram of a graph

Description automatically generatedA graph of a scatter plot

Description automatically generated**

Older people are paying higher taxes compared to younger ones also people with more BMI paying more charges not many but a notable percentage. BMI is high in older people compared to younger ones.

From 3D scatter plots we observe that there is a strong positive relationship between age and charges. There is a weak positive relationship between BMI and Age, Charges and BMI with some outliers. We cannot determine any insights with high confidence considering weak relationships.

3D scatterplot colored by different categorical variables suggest that smokers from all the regions paying higher insurance charges. There are a greater number of smokers from ‘other’ regions who pay higher insurance. Most of the non-smokers coming from southeast pay less taxes.

1. **Testing**

A one-sample Hotelling's T-squared test was conducted to assess whether the mean vector of three numeric variables—transformed charges, BMI, and transformed age—differs significantly from the population means reported in the literature. The transformed literature mean vector was [10.73, 28.5, 27.75], while the sample mean vector was found to be [10.85, 30.66, ...]. Although normality tests such as the Royston and Henze-Zirkler tests indicated potential deviations from normality (with p-values of approximately 0 and 1.845e-0, respectively), visual inspection of the plots suggested acceptable normality. Given an alpha level of 0.05, the resulting p-value from the Hotelling's T² test was 2.2e-16, which is substantially below the threshold for statistical significance. With a T² statistic of 57.357, we reject the null hypothesis and conclude that the sample mean vector significantly differs from the population mean vector reported in the literature. Therefore, the transformed population means are not a plausible match for the observed sample means.

|  |  |
| --- | --- |
| Test | p-value |
| Royston | 1.845e-0 |
| Henze-Zirkler | 0 |

A graph with a red line

Description automatically generated