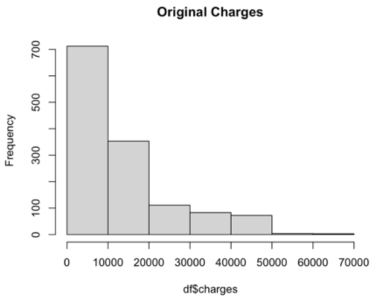
1. Univariate analysis

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'.



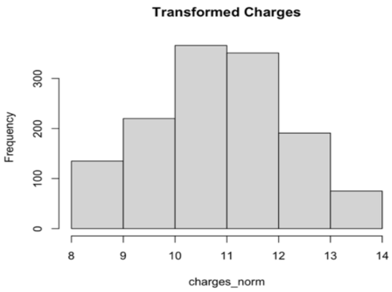


Table 1 Numerical summaries for each variable

Variables	Summaries
Charges ($, x¯ ± sd)	13270 ± 12110.01
BMI (kg/m^2, x¯± sd)	30.66 ± 6.09
Age (years, x¯± sd)	39.21 ± 14.04

2. 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.

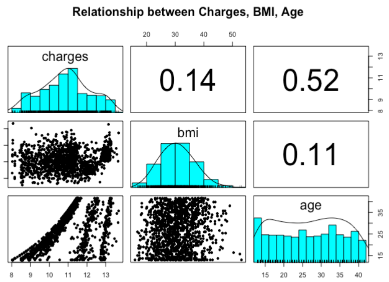
Table 2.1 Numerical summaries by ‘Smoker’

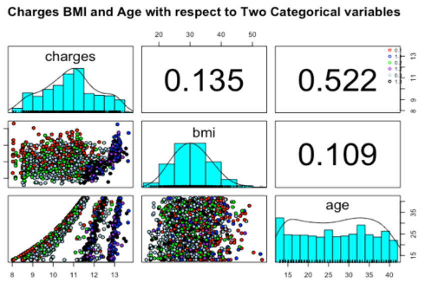
Variables	Smoking status	
	Not a smoker	Smoker
Charges ($, $\bar{x}$ ± sd)	8434 ±5994	32050±11542
BMI (kg/m^2, $\bar{x}$± sd)	30.7±6.04	30.7±6.32
Age (years, $\bar{x}$± sd)	39.4±14.1	38.5±13.9

Table 2.2 Numerical summaries by ‘Region’

Variables	Region		
	Southeast	Southwest	Other
Charges ($, $\bar{x}$ ± sd)	14735 ±13971	12347 ± 11557	12911 ± 11167
BMI kg/m^2, $\bar{x}$± sd)	33.4 ± 6.48	30.6 ± 5.69	29.2 ± 5.55
Age (years, $\bar{x}$± sd)	38.9 ± 14.2	39.5 ± 14.0	39.2 ± 14.0

3. Bivariate Analysis





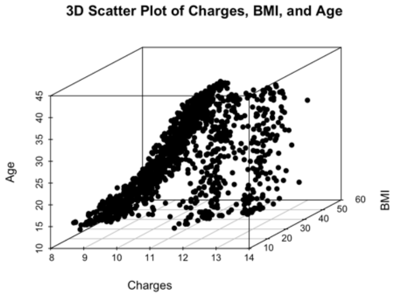
Smoker	
0	Not a smoker
1	Smoker
Region	
1	Southeast
2	Southwest
3	Other region

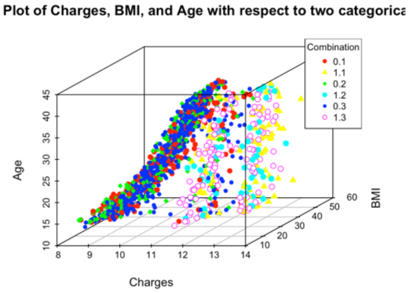
As people get older, they tend to pay more for insurance. This is because older individuals often require more medical care, leading to higher insurance costs.

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.

4. Multivariate Analysis





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.

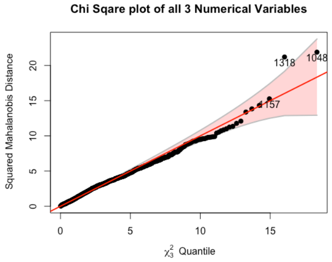
5. Testing

Sample info

N= 1338,
True mean for charges:10.73(After transformation) [Means of charges and age obtained from the literature review are transformed because we transformed the attributes charges and age in our sample]
True mean for BMI: 28.5
True mean for age: 27.75(After transformation),
Mean vector = [10.73,28.5,27.75]

Assumptions:

1. Variable: Numeric
2. Normal: Normality Acceptable (**Tests says not normal** but plot looks normal)



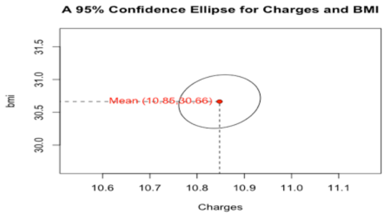
Test	p-value
Royston	1.845e-0
Henze-Zirkler	0

Hypothesis:

1. **Null Hypothesis:** Sample mean vector equal to true mean vector.
2. **Alternative Hypothesis:** Sample mean vector not equal to true mean vector.

alpha: 0.05
p-value: 2.2e-16
T-static value: 57.357
Decision: p-Val < alpha; Reject null hypothesis.
Conclusion: It appears to be mean vector of three potentially transformed numeric variables is not equal to the means obtained in the literature review, since p value (2.2e-16) obtained by hoteling T^2 test is less than the alpha (0.05), hence we reject null hypothesis.

Appropriately transformed literature means [10.73,28.5] is not a plausible value for the mean vector [10.85,30.66].



This project is entirely my work. I have not discussed about this project with anybody in or out of class. I did not use any unauthorized sources such as help from tutoring services (in person or online), nor have I used generative AI/Online assignment assisting websites (such as Chat GPT, DALL-E, Bard, Chegg, course Hero, etc.). I understand and have complied with the Rowan University's academic integrity policies outlined in (https://confluence.rowan.edu/display/POLICY/Academic+Integrity+Policy). I understand that failure to adhere to this policy will be considered as violation of academic integrity and the appropriate university procedures will be invoked (https://sites.rowan.edu/academicaffairs/academic integrity).

NAME: Dinakar Reddy Donthireddy

DATE: 03/18/2024

SIGNATURE: Dinakar Reddy Donthireddy