Exploratory Data Analysis of Hospital Charges for Insurance Pricing Strategy

Abstract:

The project aims to conduct an Exploratory Data Analysis (EDA) to scrutinize the hospital charges associated with patient care within a specified dataset. This comprehensive analysis will delve into identifying patterns and relationships between patient demographics and the resultant hospital charges. Through rigorous examination of these variables, the study seeks to unearth insights that could serve as a foundational base for devising strategic recommendations for insurance fee structuring. The ultimate goal of this project is to leverage statistical and analytical methodologies to offer data-driven suggestions that could optimize insurance pricing models, ensuring they are both equitable for patients and financially viable for insurance providers.

Explore Data Analysis:

The insurance dataset contains 6 features, including 3 numerical and 3 nominal features, and one target variable “charges” with a total of 1338 observations. There are no missing values in this dataset. The R output of the data structure, described statistic and missing value statistics are provided below:

Data Structure:

```
'data.frame':    1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : chr  "female" "male" "male" "male" ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : chr  "yes" "no" "no" "no" ...
 $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

Describe Statistic:

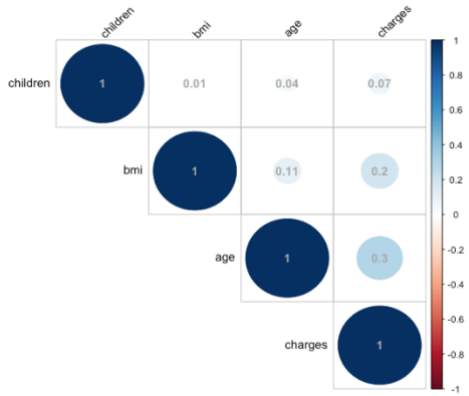
age	sex	bmi	children	smoker	region	charges
Min. :18.00	Length:1338	Min. :15.96	Min. :0.000	Length:1338	Length:1338	Min. : 1122
1st Qu.:27.00	Class :character	1st Qu.:26.30	1st Qu.:0.000	Class :character	Class :character	1st Qu.: 4740
Median :39.00	Mode :character	Median :30.40	Median :1.000	Mode :character	Mode :character	Median : 9382
Mean :39.21		Mean :30.66	Mean :1.095			Mean :13270
3rd Qu.:51.00		3rd Qu.:34.69	3rd Qu.:2.000			3rd Qu.:16640
Max. :64.00		Max. :53.13	Max. :5.000			Max. :63770

Missing Value Statistics:

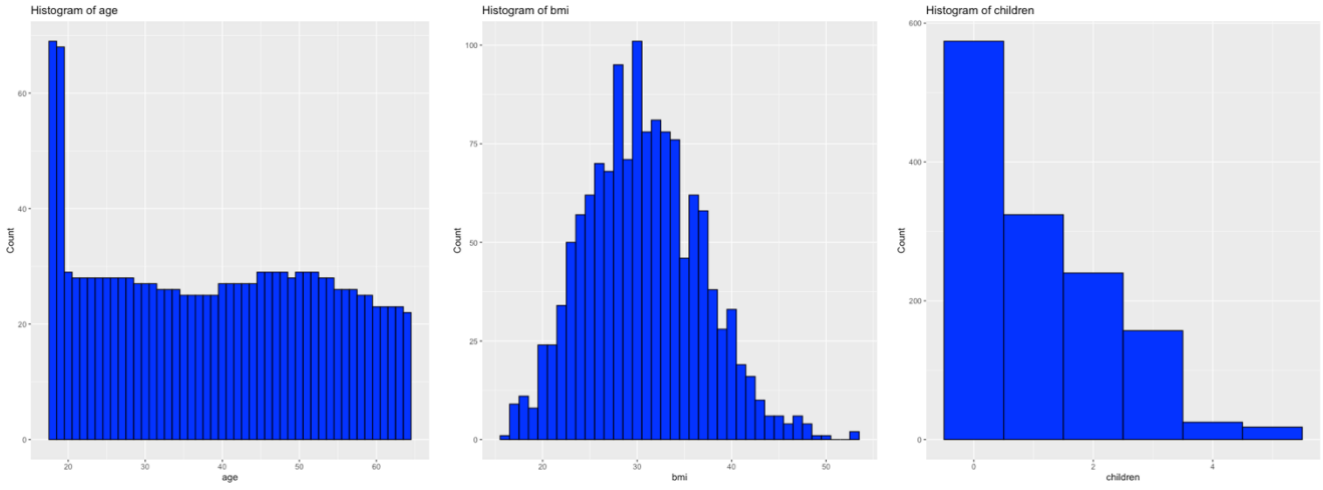
age	sex	bmi	children	smoker	region	charges
0	0	0	0	0	0	0

Correlation Matrix of Numerical Features:

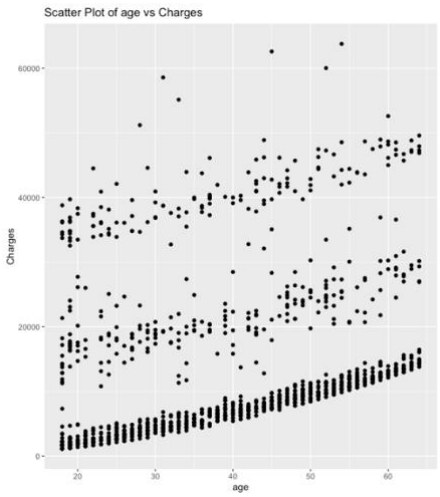
The correlation matrix shows that numerical features only have a weak linear correlation with the target variable, charges.

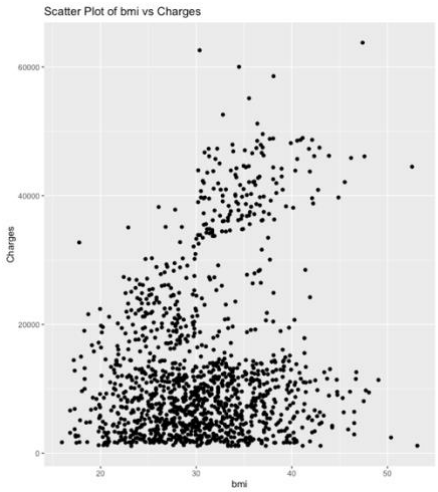


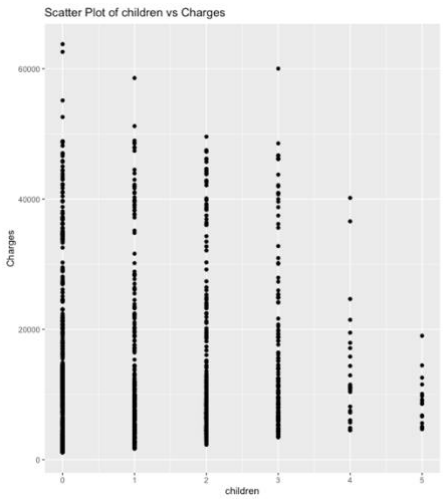
Distribution of Numeric Features:
The age distribution is equal except for a spike under 20.
The BMI distribution has a spike of around 30.
The number of children distribution has a decreasing trend with the number of children increasing.



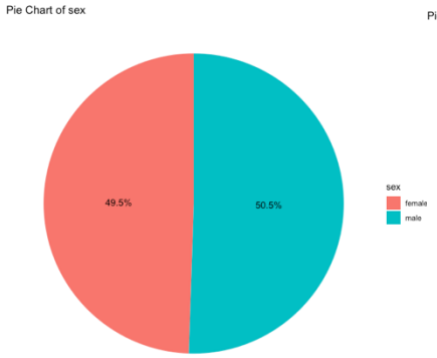
Scatter Plot of numerical feature V.S. Charges:
Age V.S. Charges: The lowest charging price increases while the age increases.
BMI V.S. Charges: There is no significant pattern in the scatter plot of BMI V.S. Charges.
Children V.S. Charges: The Price range becomes shorter with the increasing number of children.

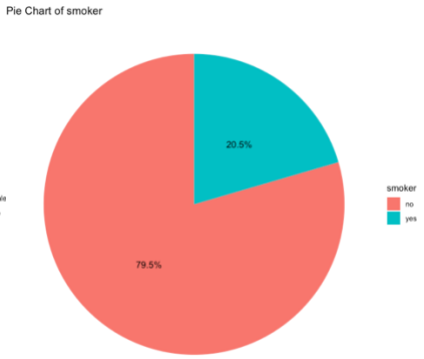


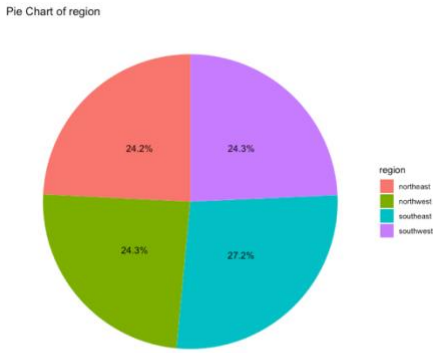




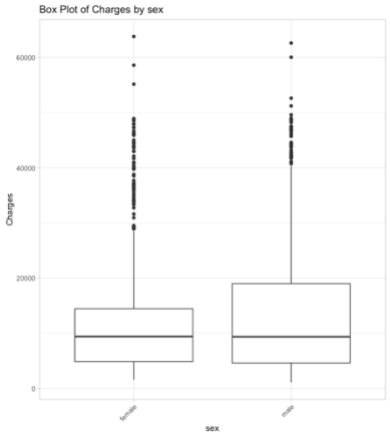
Pie Chart of Nominal Data:
Sex: The ratio of the sex is almost equal in the dataset.
Smoker: The ratio of the non-smoker is more than smokers in the dataset.
Region: The ratio of the region is almost equal in the dataset.

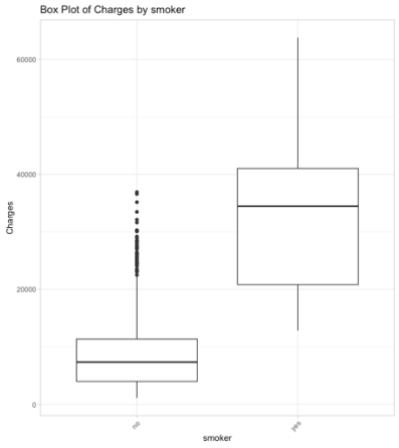


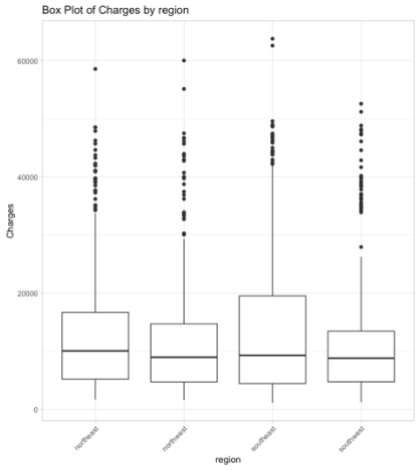




Box plot of nominal data:
Sex V.S. Charges: The median is similar between genders but the male has a wider interquartile range(IQR) which indicates there is a similar population of charges between genders but a higher variability of charge in males.
Smoker V.S. Charges: The median and whisker of smokers are significantly higher and longer than non-smokers indicating that smokers have higher charges and a larger range of charge than non-smokers.
Region V.S. Charges: Boxplots of four regions have similar median and slight differences in interquartile range(IQR). It shows that the region doesn't affect charges significantly.







Conclusion:

According to the exploratory data analysis, we find out that higher age and smoking people will have higher amounts of hospital charges. Therefore, the insurance company could raise the insurance fee for smoking people and elder people. Moreover, we discovered that male population and people who have fewer children have a higher range of hospital charges, so the company should develop a more detailed analysis of these two populations to distinguish more segments for pricing in these two groups.