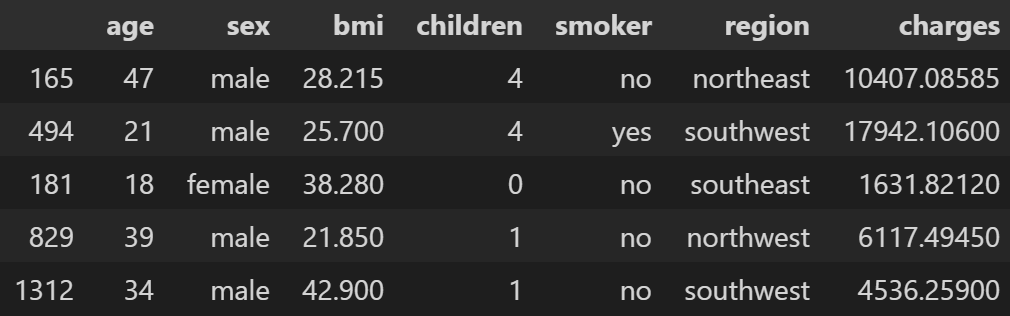
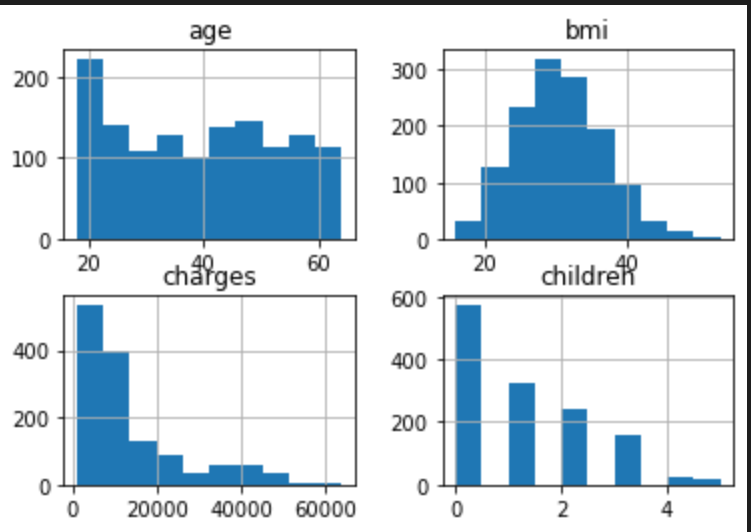
Below is a 5 random sample data set:

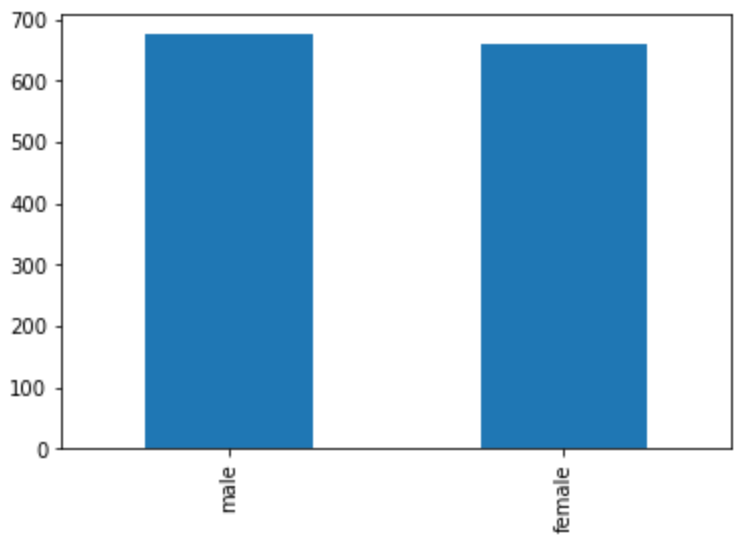
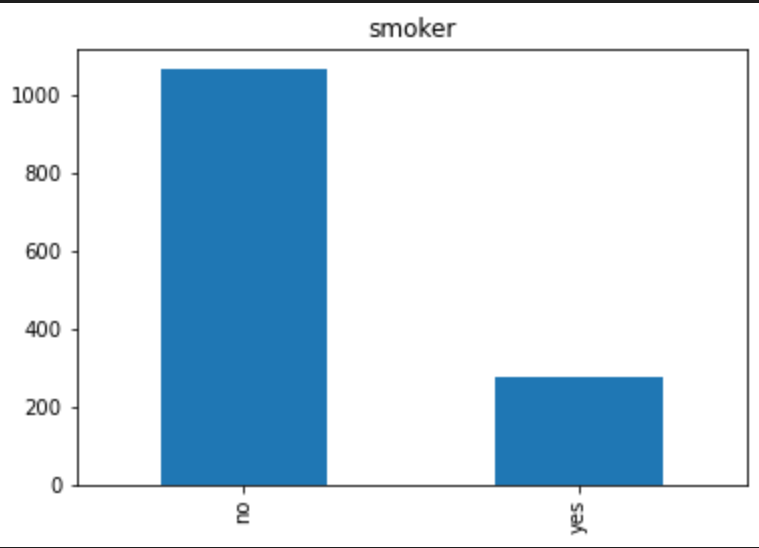
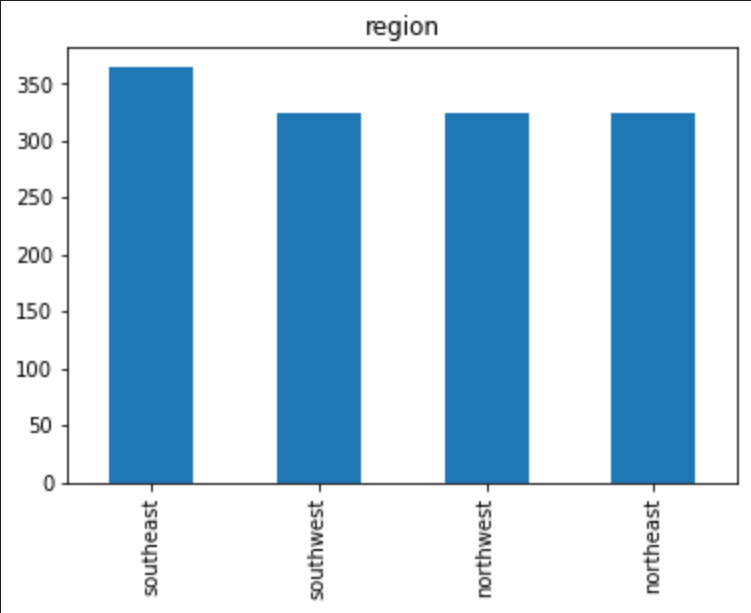


For my final project, I am going to analyze the data of personal medical costs across the United States. I downloaded the data from the website “Kaggle.com”. From the sample dataset above, we can see that it has 7 columns which are age, sex, bmi, children, smoker, region, and charges. Besides, we know that the categorical variables are sex, smoker, and region. The numeric variables are age, children, bmi, and charges. I will focus on relationships between the column variables and the charges which is the medical cost.



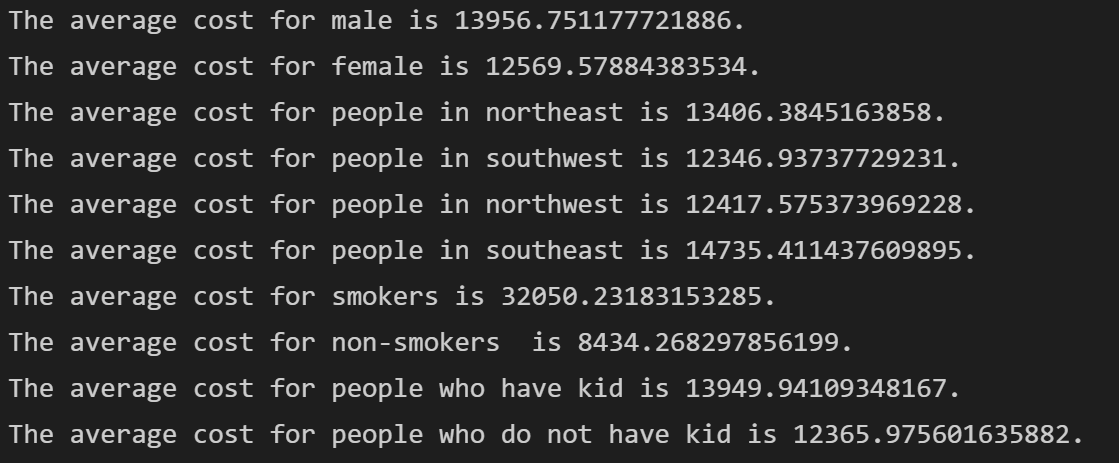
From the description above, we know that the average medical cost overall is about 13270, and the minimum cost is about 1122, the maximum cost is about 63770.



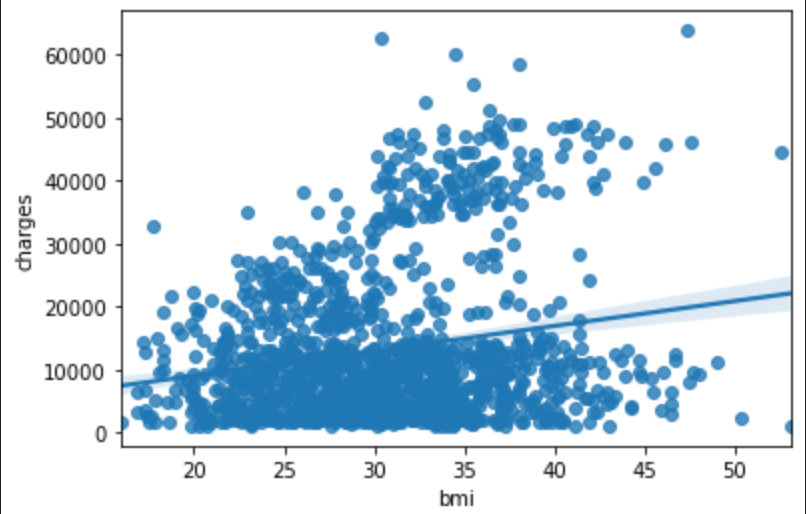
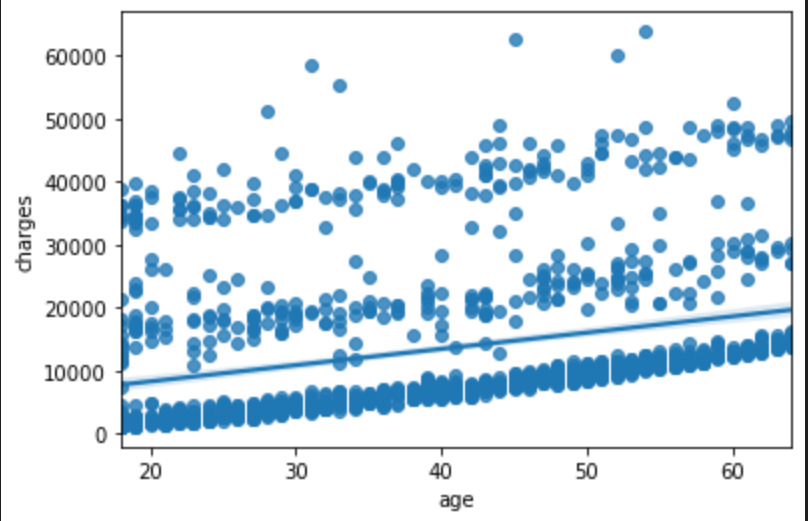
From the histograms and bar plots above, we can have a general idea of the value count of each numeric variable. We can see that the histogram of bmi approximately follow a normal distribution.

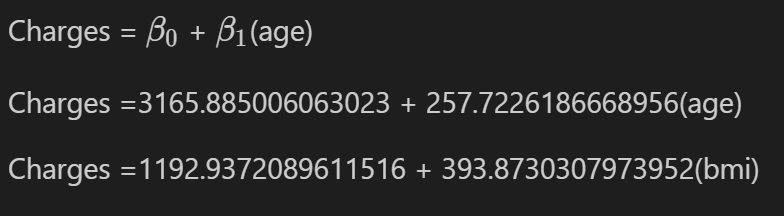
**SEX&SMOKER#REGION:**



Based on the average cost that I calculated above, I can conclude that males spend more money than females on medical things, and people who live in the southeast seem to pay more than the other areas, people in southwest pay the less on average. Also, people who smoke pay a lot more than those who do not smoke, so it indicates that smokers are much more easy to get sick. Besides that, the figure above also mentions that people who have kids are paying more than people who do not have kids, but the gap in cost is not so big.

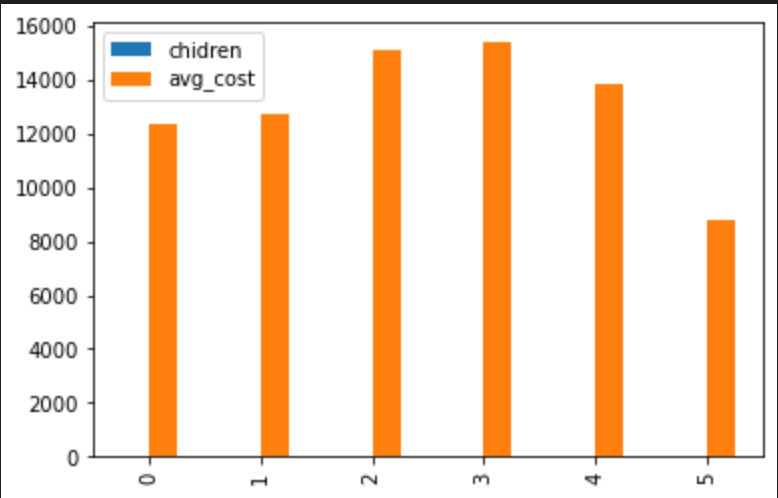
**AGE&BMI:**





From the scatter plots above, we can see that they both have positive relationships between age and charges and bmi and charges. This means that the medical cost will increase while either the values of age or bmi increase. Based on the figure, I can tell that the bmi has a bigger impact on medical most than age because the slope of the sample linear regression to bmi is slightly larger than the slope of the sample linear regression to age; besides, we can see that the visual line of bmi in the scatters plot is a little more oblique than the visual SLR of age.

**CHILDREN:**

Based on the figures above, I can conclude that, on average, people who have 3 kids cost the most on personal medical, and people who have 5 kids spend less on personal medical compared to others.

**SUMMARY:**

To sum up, based on the analysis above, we know that the average medical cost overall is about 13270, and the minimum cost is about 1122, the maximum cost is about 63770. I can conclude that all those variables matter on the medical cost in the United States at a different level. In age and bmi, the charges of medical increase while they increase. Besides, males, people in the southeast, smokers, and people who have kids are most likely to pay more for their medical things. By the factor children, more specifically, the figures tell that people who have 5 kids are spending even less than those people who do not have kids, on average, and people who have 3 kids spend the most on medical stuff. Also, I found two SLR between age and bmi to charges. In conclusion, people who smoke, people who are getting older, people who have larger bmi, people who have 3 kids, people who are male, and people who live in the southeast are most likely charges more in the United States.