Healthcare Insurance Analysis of Predictor Factors of Medical Charges

Questions explored:

- What is the average age of the insured population?
    - 39 years
    - Is there a correlation between age and insurance charges?
        - There is a weak correlation (0.3) between age and charges
    - For the analysis we subcategorize the group into:
        - young adults, for age < 35
        - middle-aged adult, for age <50 and >= 35
        - mature adult, for age <65 and >=50
        - senior, for age>=65

- What is the distribution of BMI across different age groups?
    - To answer this question, we categorized our group into:
        - underweight, if BMI <18.5
        - healthy weight, if BMI >= 18,5 and <25
        - overweight, if BMI >= 25 and <30
        - obese, if BMI >= 30
    - The distribution is fairly uniform. All group ages display a predominance of obesity, followed by overweight individuals.
    - Young adults have 8% more people with a healthy weight than mature adults.
    - The rate of healthy-weighted people among young adults is approximately 66% higher than the rate among mature adults. (20% x 12%)
    - Is there a correlation between BMI and insurance charges?
        - There is a weak correlation (0.2) between BMI and charges

- How does the number of children correlate with insurance charges?
    - There is a very weak correlation between these two variables (0.067)

- What is the distribution of insurance charges by region?
    - The 4 regions present similar medians of medical charges
    - There are outliers in each region, indicating some individuals with significantly higher charges - smokers?
    - Southeast has the highest maximum values, excluding outliers
    - Southwest has the narrower distribution of medical charges

- How does smoking status impact insurance charges?
    - Smoking is related to higher charges across people with the same BMI, same age group and in every region.
    - The increase in charges seem to be more pronounced the higher the BMI

- How do insurance charges differ between males and females?

    - There is only a small difference between the mean charges of male and females, with males having charges 11% higher than females. But, males also smoke more than females (The percentage of male smokers is 6.3% higher than the percentage of female smokers.), which could be the real reason why males have slightly higher medical charges than females.
    - 20.4% overall smoking rate
    - 17.3% females are smokers
    - 23.5% males are smokers

- Is there a significant difference in insurance charges between smokers and non-smokers?
    - smokers have charges 3.7 times higher than non smokers on average (not isolated by BMI category)
        - underweight: smokers' charges are 3.4 times higher than that of non smokers
        - healthy weight: smokers' charges are 2.59 times higher than that of non smokers
        - overweight: smokers' charges are 2.72 times higher than that of non smokers
        - obese: smokers' charges are 4.7 times higher than that of non smokers

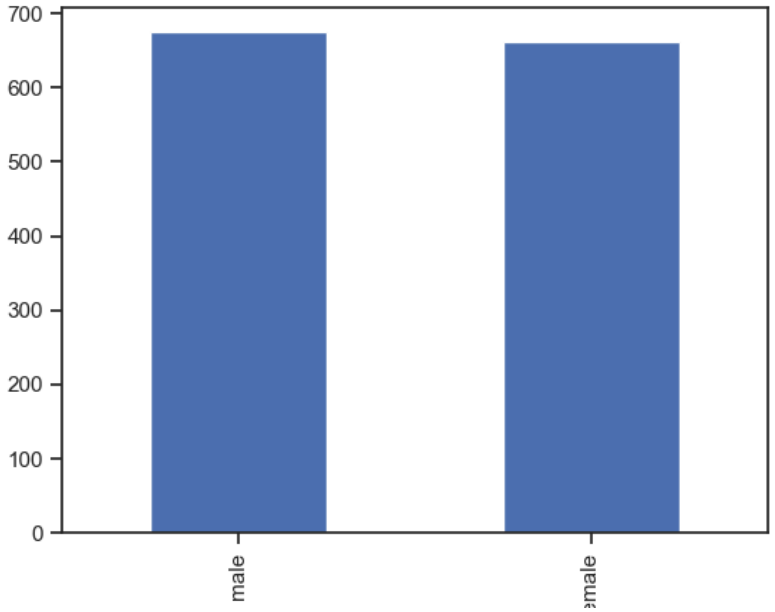
Smoking is the single most important predicting factor of high medical charges.

Predictive Modeling:

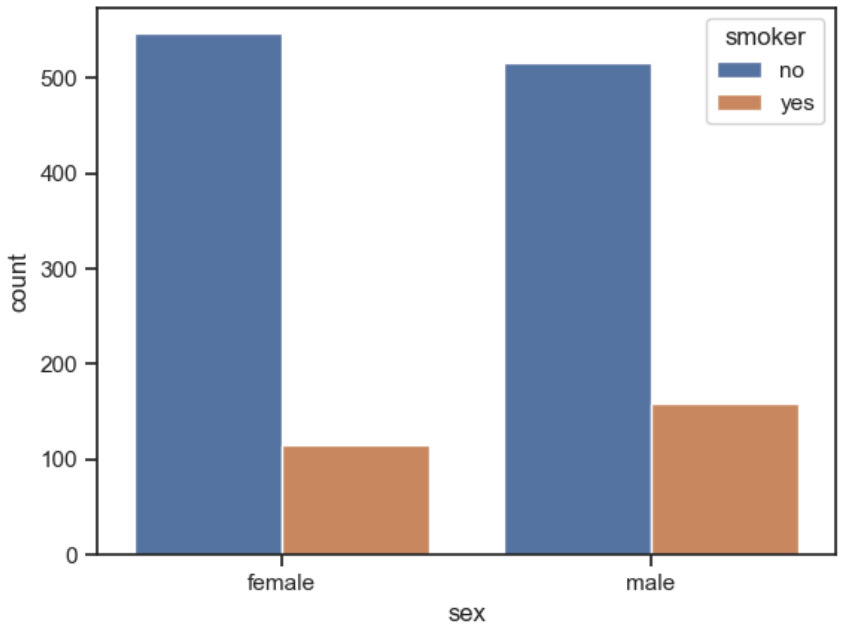
The three most significant predictors of high medical charges are smoking, obesity, and age over 42. The complex interplay between these factors suggests that random forest modeling would be suitable for the next phase of this project.

Supporting Viz:

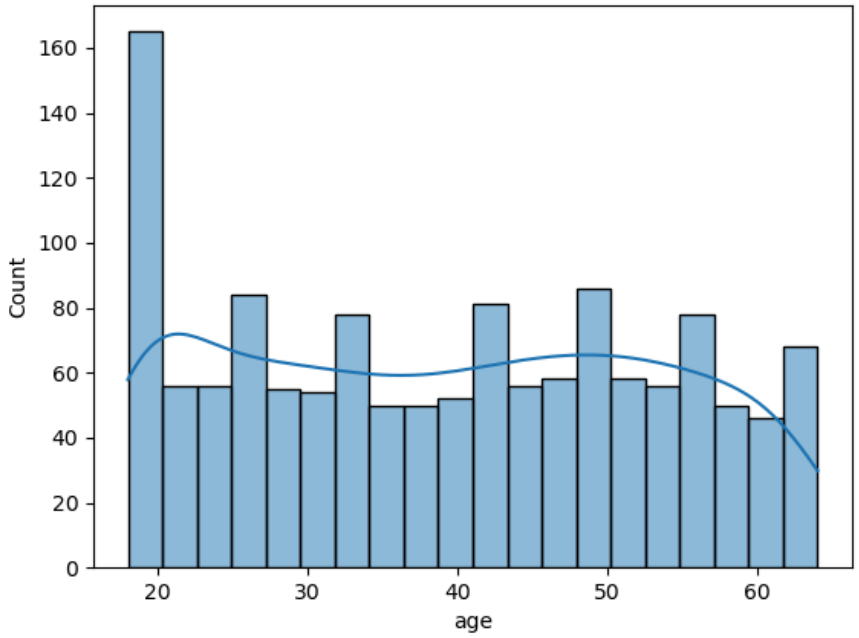
Sex Distribution



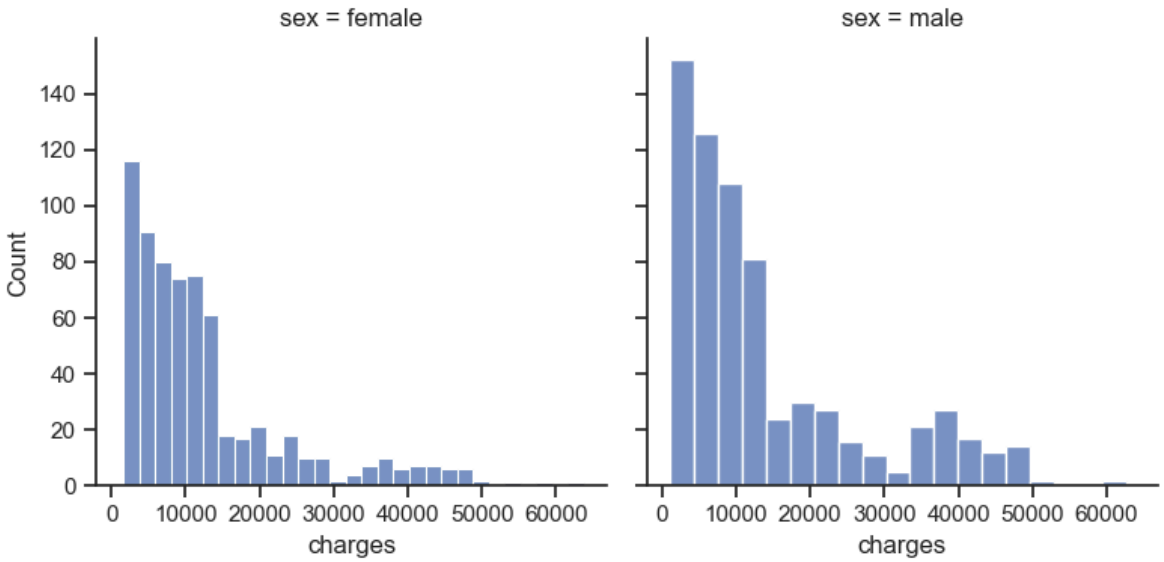
Smoking Distribution by Sex



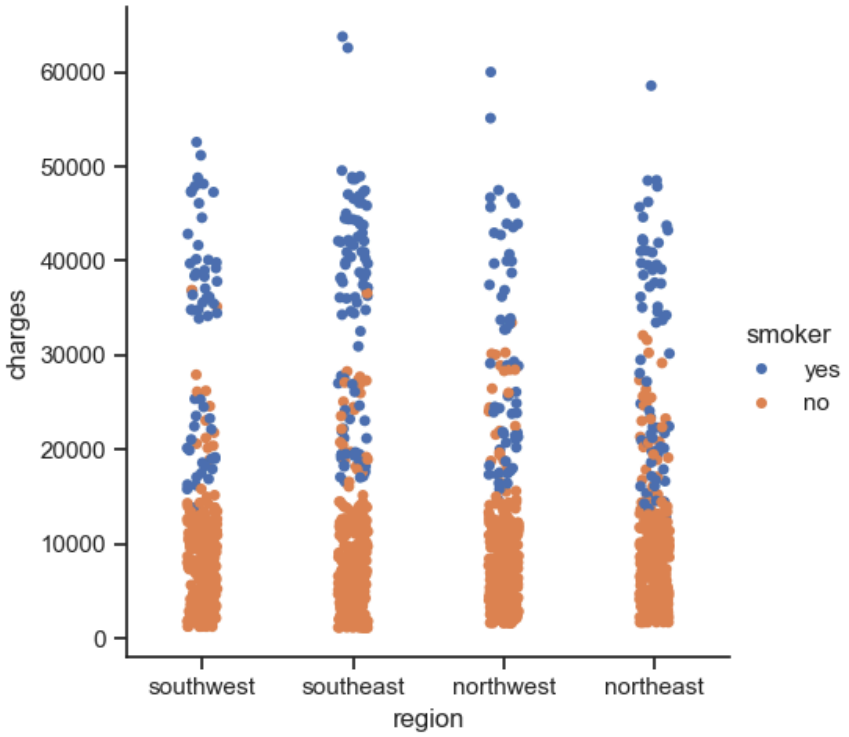
Age Distribution of population



Distribution of Charges by Sex



Medical Charges by Region - separated by smokers and non-smokers



Pairplots of age, bmi, number of children and medical charges: