**Project Report**

Sahil Agarwal

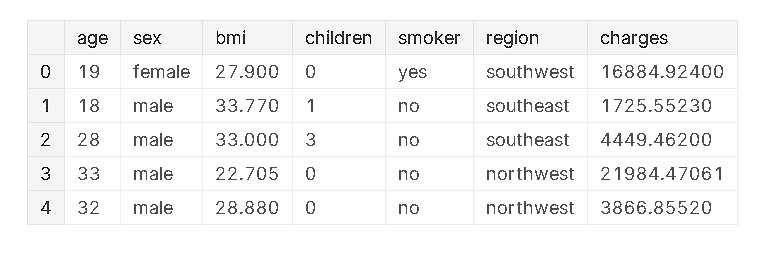
(20CS3054)

**Introduction:**

The goal of this project is to explore a dataset that provides information about the cost of treatment of different patients. The dataset contains various features such as age, sex, bmi, smoker or not, region, and the cost of treatment. The aim of this project is to understand how different features relate to the cost of treatment and to practice regression analysis.

**Data Cleaning:**

Before starting the analysis, the data set was first checked for any missing values. It was observed that there were no missing values, which is always a good sign when working with datasets.

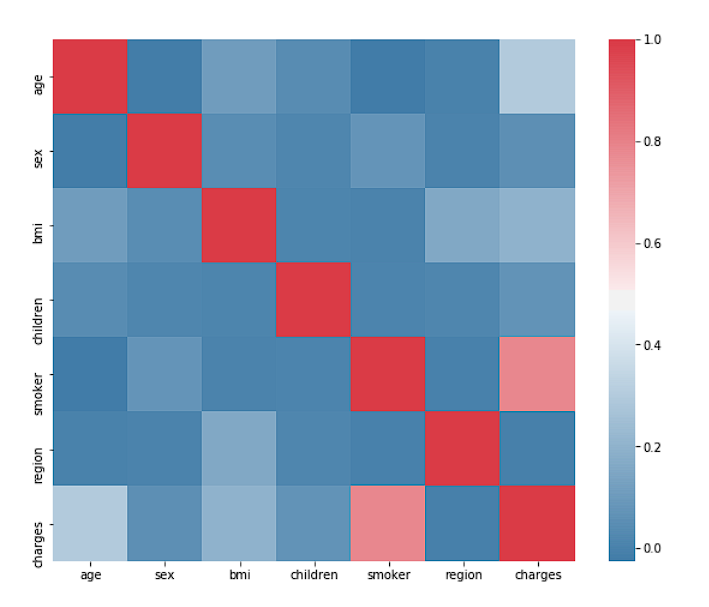


**Encoding Categorical Features:**

Categorical features such as sex, smoker or not, and region were encoded using the LabelEncoder. While OneHotEncoder could have been used for encoding categorical variables with large variability, in this case, since there is no special order in which the regions would be listed, LabelEncoder was used instead.

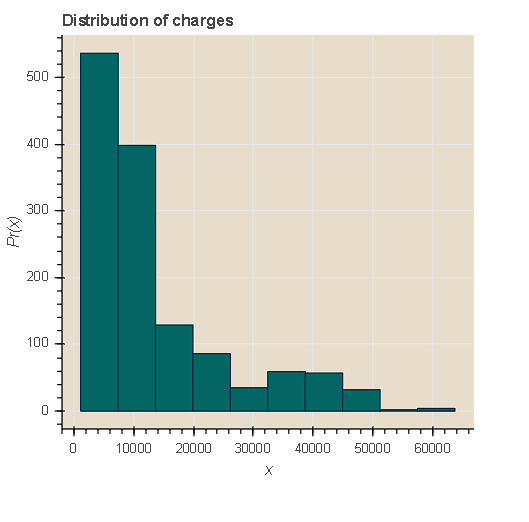
**Correlation Analysis:**

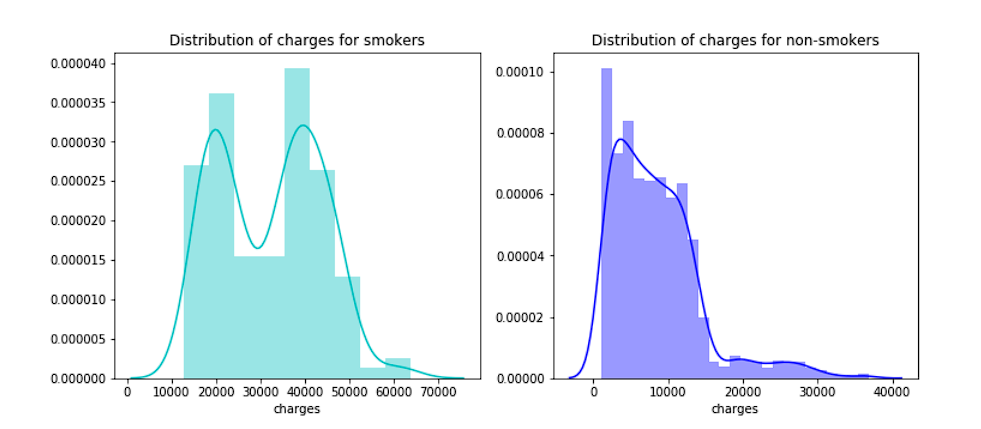
A correlation analysis was performed to understand which features are more strongly correlated with the cost of treatment. It was found that the smoking feature was the only feature that had a strong correlation with the cost of treatment. Interestingly, the BMI feature did not show a strong correlation with the cost of treatment.



**Data Visualization:**

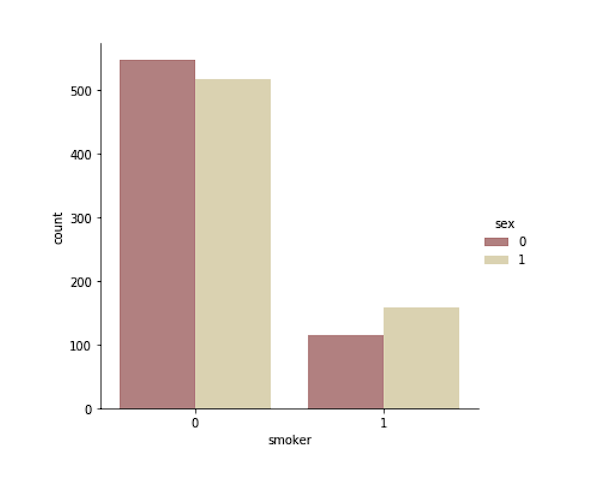
Several visualizations were created to better understand the relationships between different features and the cost of treatment. A heatmap was used to visualize the correlations between different features. Bokeh was used to create an interactive histogram showing the distribution of charges. Seaborn was used to create a distribution plot for charges of smoking patients, and a scatter plot to visualize the relationship between age and charges.

****

****

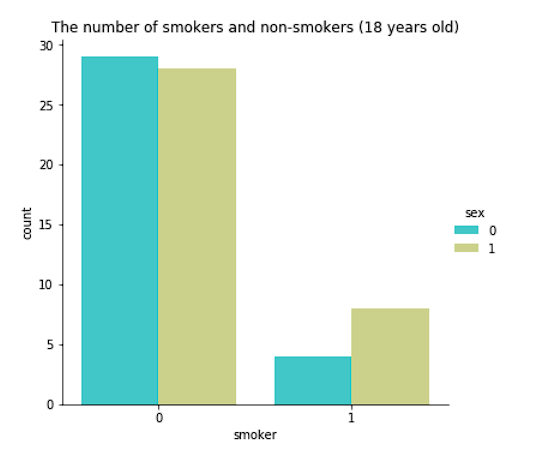
**Analysis of Smoking Patients:**

It was observed that smoking patients spend more on treatment. Additionally, it was noted that there are more male smokers than female smokers in the dataset. It was assumed that the total cost of treatment in men would be more than in women, given the impact of smoking. However, this was not confirmed in the analysis.



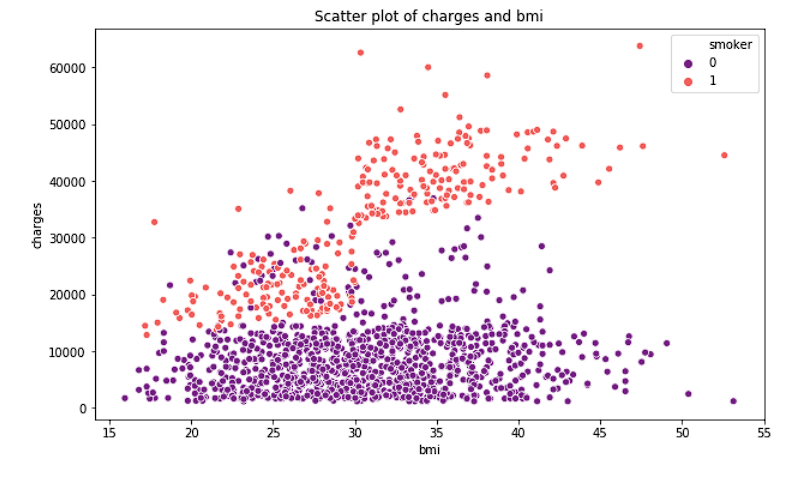
**Analysis of Age and Cost of Treatment:**

The analysis of age and cost of treatment revealed that in non-smokers, the cost of treatment increases with age. This makes sense as age is generally associated with an increased risk of health issues. However, in smoking patients, this relationship is not observed. It was noted that there are patients under 20 years of age in the dataset, with the youngest patient being 18 years old. Smoking patients at the age of 18 were found to spend much more on treatment than non-smokers.

****

**Analysis of BMI and Cost of Treatment:**

While it was surprising that BMI did not show a strong correlation with the cost of treatment, it was observed that patients with a BMI greater than 30 do tend to spend more on treatment than those with a BMI less than 30. This is expected as obesity is generally associated with a higher risk of health issues.

****

**Conclusion:**

In conclusion, the analysis of the dataset revealed interesting insights into the factors that affect the cost of treatment. It was found that smoking is the strongest predictor of the cost of treatment, and patients with a BMI greater than 30 tend to spend more on treatment. The analysis also revealed that non-smoking patients tend to spend more on treatment as they age, while this relationship is not observed in smoking patients. However, it should be noted that the dataset used in this analysis is relatively small, and a larger dataset would provide more reliable results. Overall, this project serves as a good example of how data analysis can provide useful insights into various aspects of healthcare. 