Sheikh-Sedat Touray
DSP562
Final Project
Spring 2024

Health Insurance Cost Analysis

Table of Contents:

1. Abstract

This project is to perform an analysis of a comprehensive Insurance dataset containing personal attributes such as age, sex, BMI, family size, and smoking habits, along with geographic factors, to understand their impact on medical insurance charges. Through exploratory data analysis and predictive modeling, insights will be gained into the key drivers of insurance costs and potential avenues for optimizing healthcare expenses. The findings will highlight the significance of certain variables in predicting insurance charges, providing valuable insights for insurance companies and healthcare providers to better tailor their services and pricing strategies.

2. Introduction

The particular research question in this study revolves around understanding the relationship between personal attributes, geographic factors, and medical insurance charges. This inquiry is of paramount importance in the field of healthcare economics and insurance pricing, as it sheds light on the determinants of healthcare expenses and aids in the development of more accurate predictive models. The increasing costs of healthcare services and the need for efficient resource allocation make this question particularly pertinent. The decision to work on this question stems from the recognition of the significant impact that insurance charges have on individuals, families, and healthcare systems, as well as the potential to inform policy decisions and improve healthcare affordability and accessibility through data-driven insights.

Below shows the first five rows and all seven columns. This dataset has 1338 rows and seven columns.

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

2.1 Problem Formulation and Literature Review

Medical insurance charges are influenced by a myriad of factors, including personal attributes and geographic considerations. Previous research has extensively explored the relationship between these variables and healthcare costs to better understand healthcare expenditure patterns and inform policy decisions.

Several studies have highlighted the significance of demographic factors such as age, gender, and family size in predicting medical insurance charges. For instance, older individuals tend to incur higher healthcare costs due to increased healthcare utilization and prevalence of chronic conditions. Gender-based differences have also been observed, with some studies indicating that women generally have higher healthcare costs than men, attributed to factors such as reproductive health needs and longer life expectancy.

Moreover, lifestyle factors like smoking habits and body mass index (BMI) have been identified as significant predictors of medical insurance charges. Smokers tend to have higher healthcare costs due to the increased risk of developing smoking-related illnesses, while individuals with higher BMIs are more prone to chronic conditions like diabetes and heart disease, leading to elevated healthcare expenses.

Geographic factors, including regional variations in healthcare infrastructure, provider density, and cost of living, also play a crucial role in determining medical insurance charges. Studies have demonstrated that individuals residing in urban areas often face higher healthcare costs compared to their rural counterparts, attributed to the availability of specialized healthcare services and higher healthcare resource utilization rates in urban settings.

Despite the extensive research in this area, there remains a need for comprehensive analyses that integrate multiple personal attributes and geographic factors to develop robust predictive models for estimating medical insurance charges. This study aims to address this gap by leveraging a rich dataset encompassing a wide range of variables to provide insights into the complex interplay between individual characteristics, geographic influences, and healthcare expenses. Some links below for more information on this:

https://www.experian.com/blogs/ask-experian/factors-that-affect-health-insurance-premium-costs/

https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance/data

Data Description:

This dataset was collected from Kaggle and it has 1338 rows and seven columns and the variables are described below.

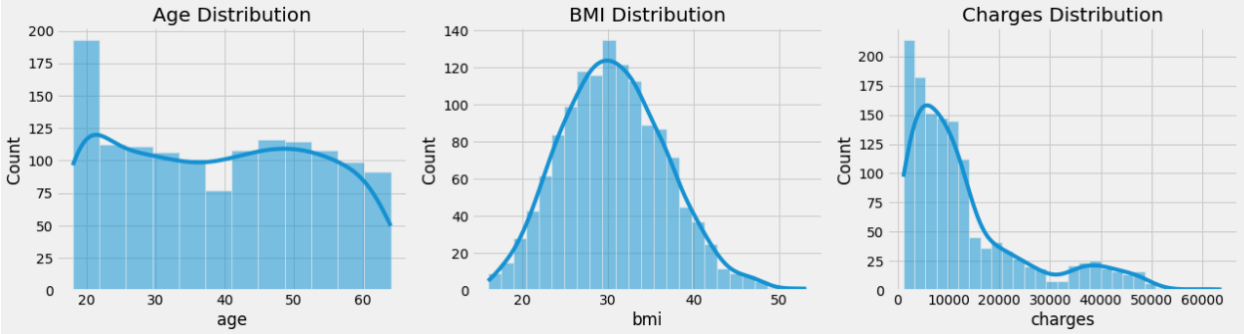
- **Age**: The insured person's age.
- **Sex**: Gender (male or female) of the insured.
- **BMI** (Body Mass Index): A measure of body fat based on height and weight.
- **Children**: The number of dependents covered.
- **Smoker**: Whether the insured is a smoker (yes or no).
- **Region**: The geographic area of coverage.
- **Charges**: The medical insurance costs incurred by the insured person.

3. Data Characterization, Descriptive Analysis, and Visualization

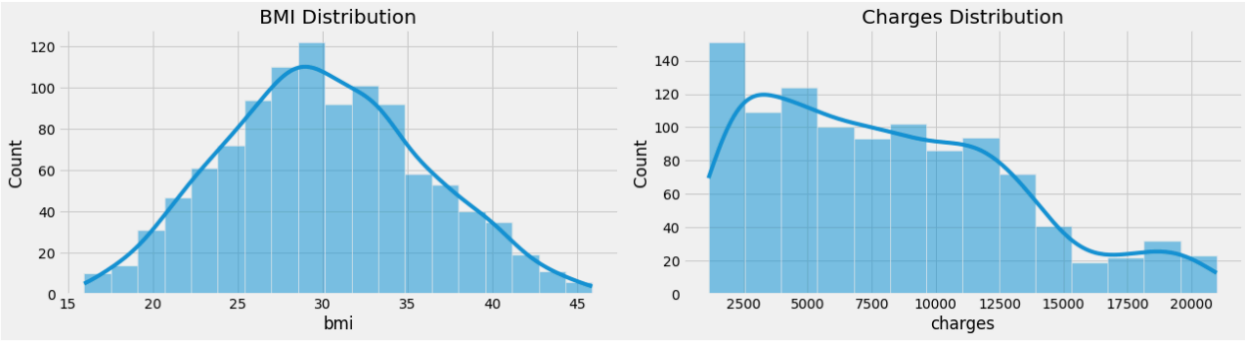
3.1 Data Issues

During the initial data exploration phase, several data issues were looked into, including missing and NAN values, outliers, and skewed distributions. Outliers, defined as observations that deviate significantly from the rest of the data, were detected using graphical methods such as histograms. Skewed distributions were observed in certain variables, indicating the need for data transformation techniques to achieve normality.

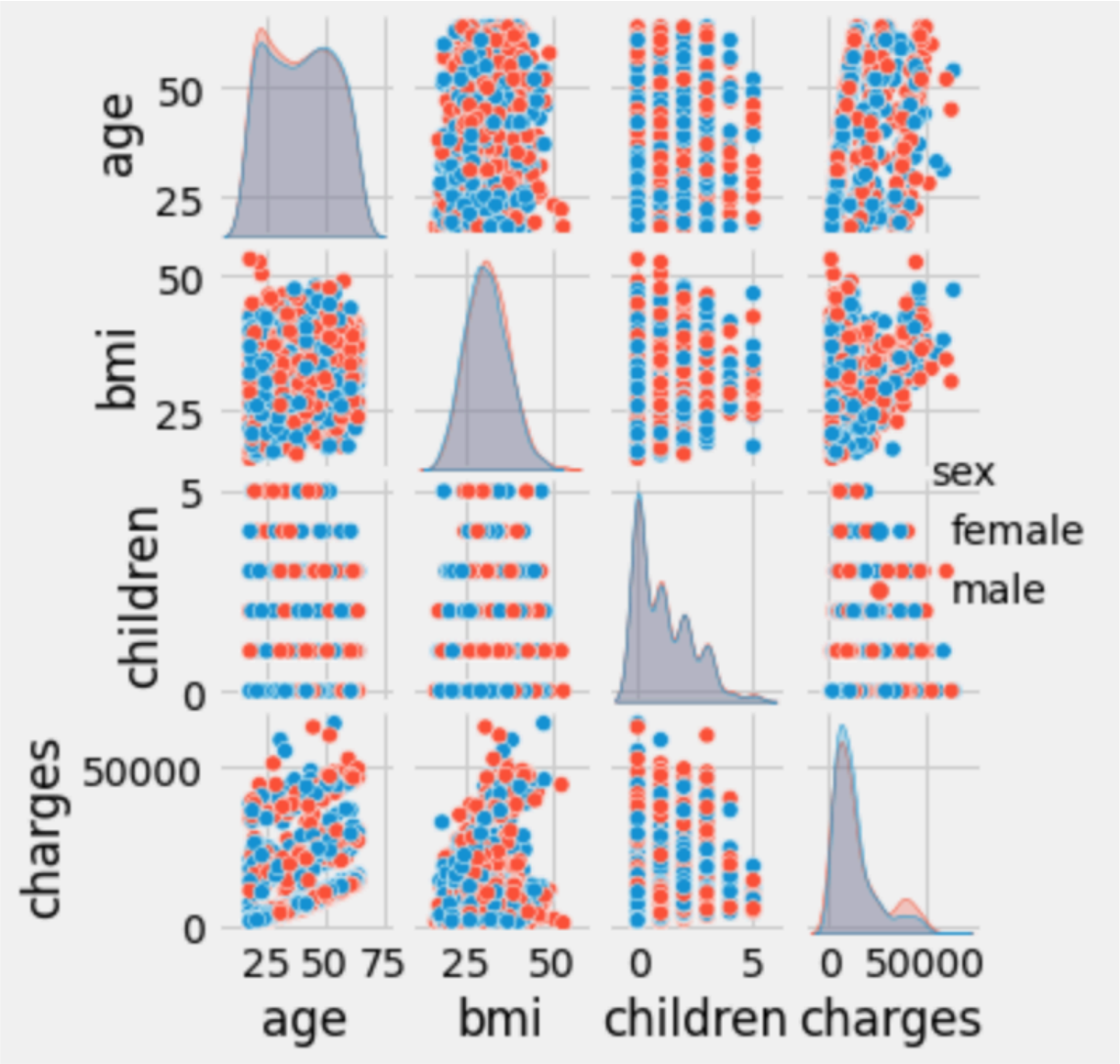
In the histograms below we can observe that distribution age seemed to be more normally distributed with a higher frequency of younger individuals, BMI appears skewed to the right and centered around the 30 mark, and charges are skewed to the right, indicating a higher frequency of lower charges with fewer instances of very high charges.



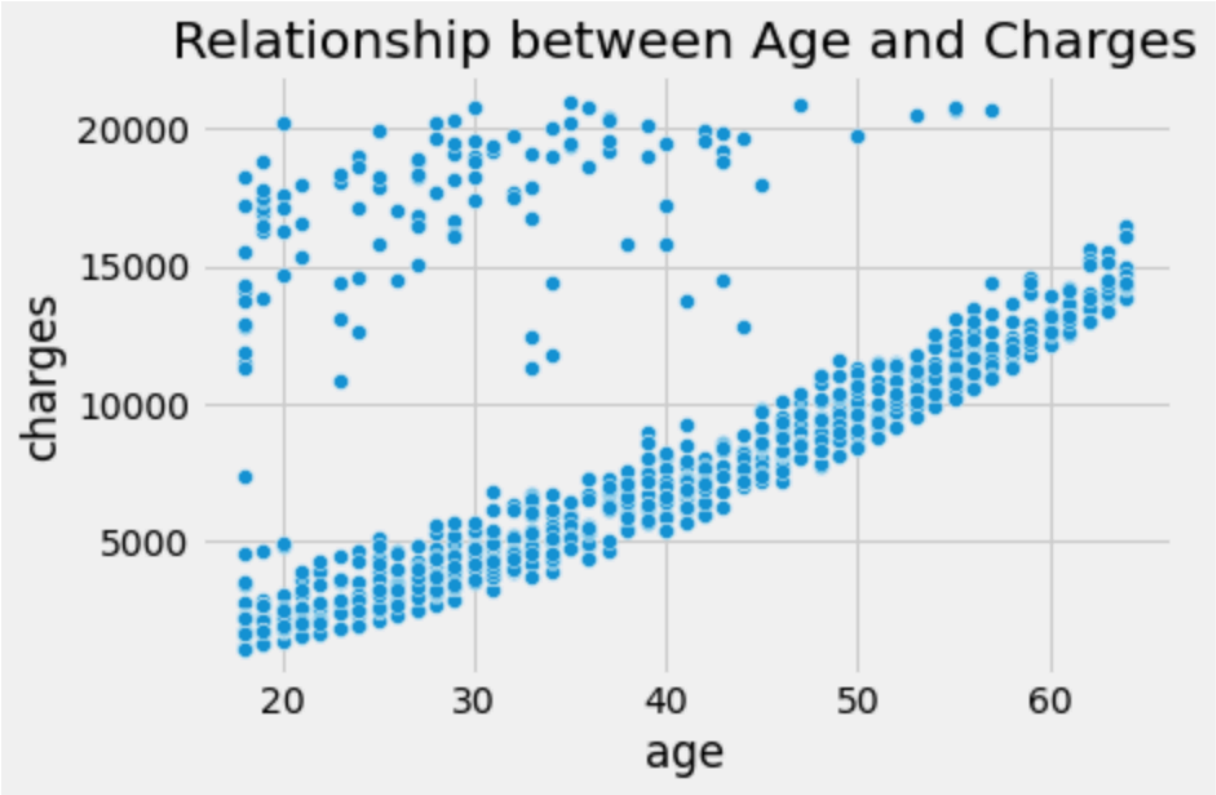
Handling outliers: For charges and BMI I defined a threshold to remove the outliers for BMI I determined 46 as the threshold $21000 for charges. And as can be seen below the distributions seem to be more normal.



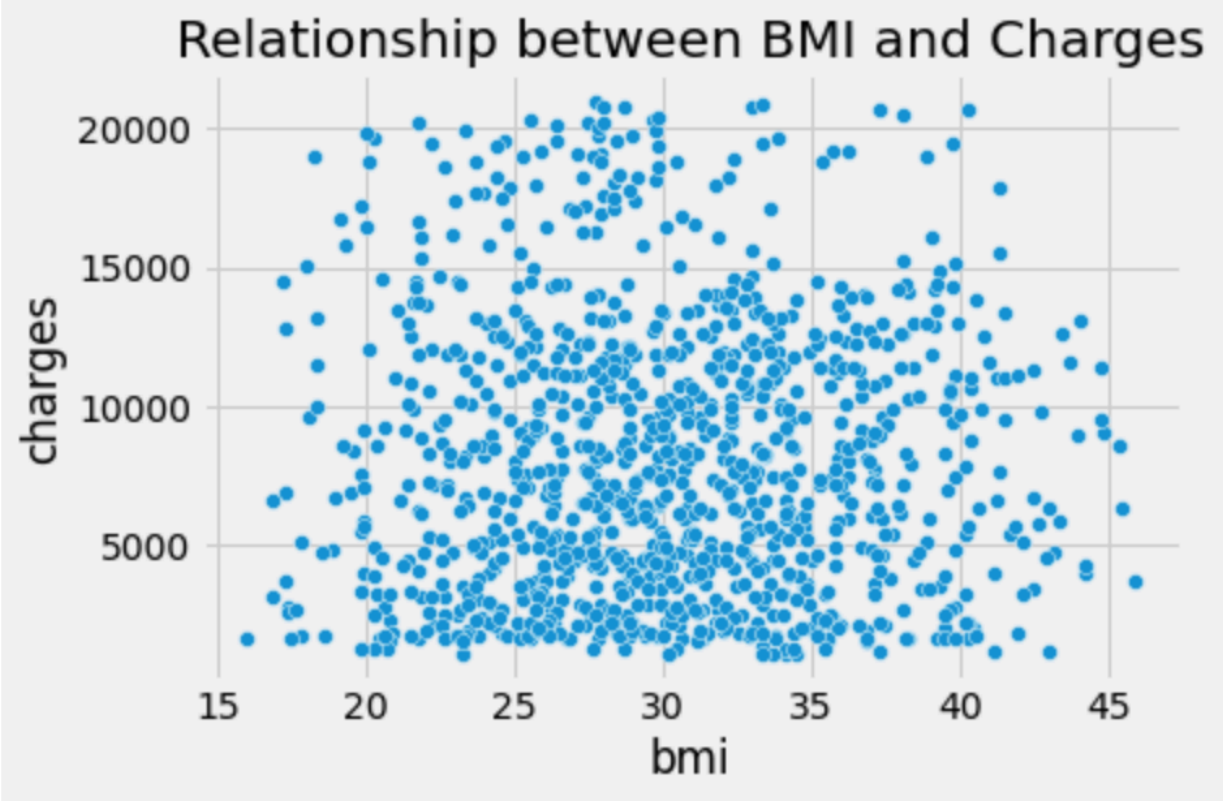
In the pairplot below we see that there is a correlation between age and charges and somewhat of a correlation between bmi and charges before the outliers were removed.



After the outliers were removed we observed the linear relationship between age and charges and BMI and charges .

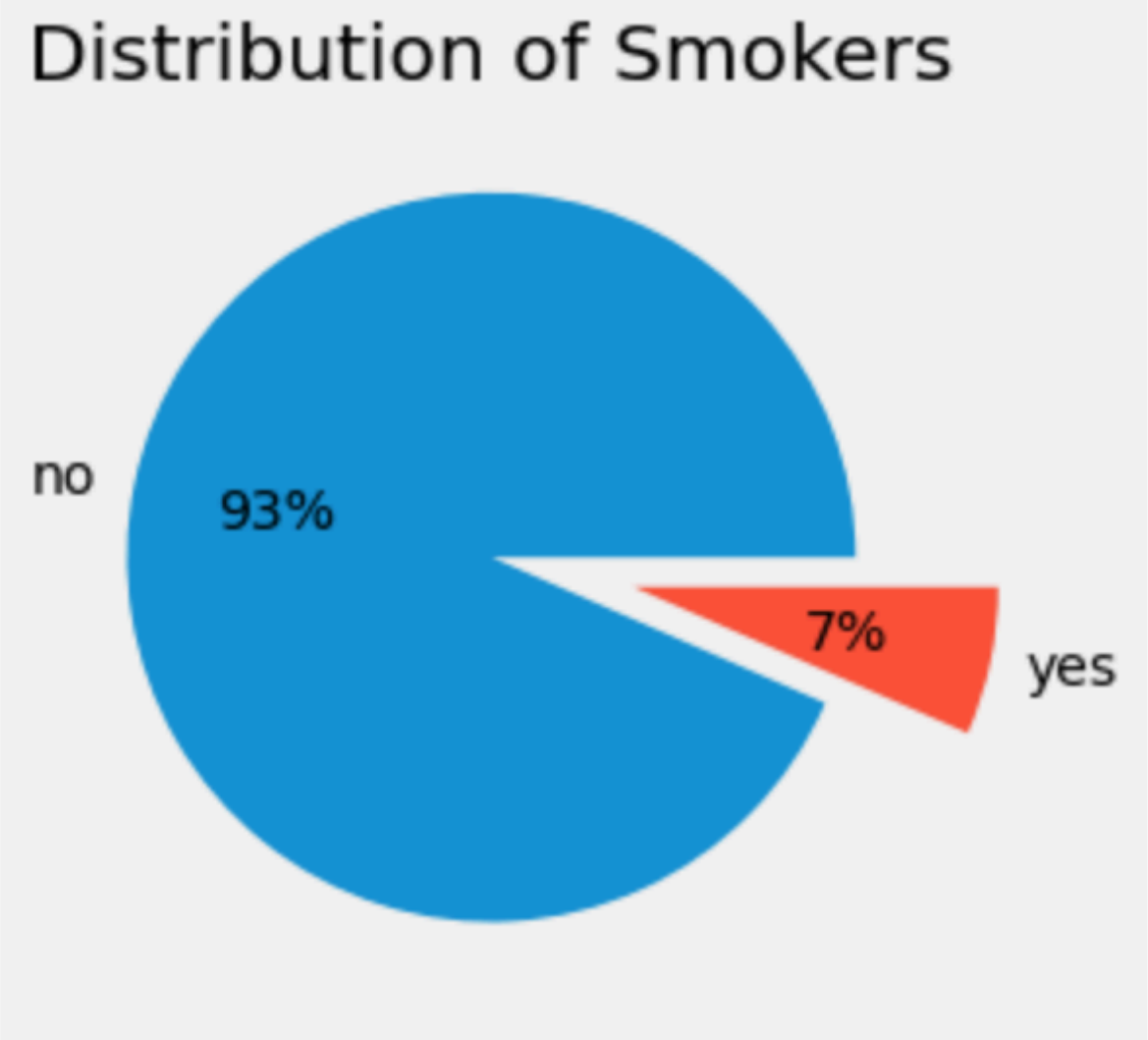


As can be seen above the correlation between age and charges is a strong positive relationship.

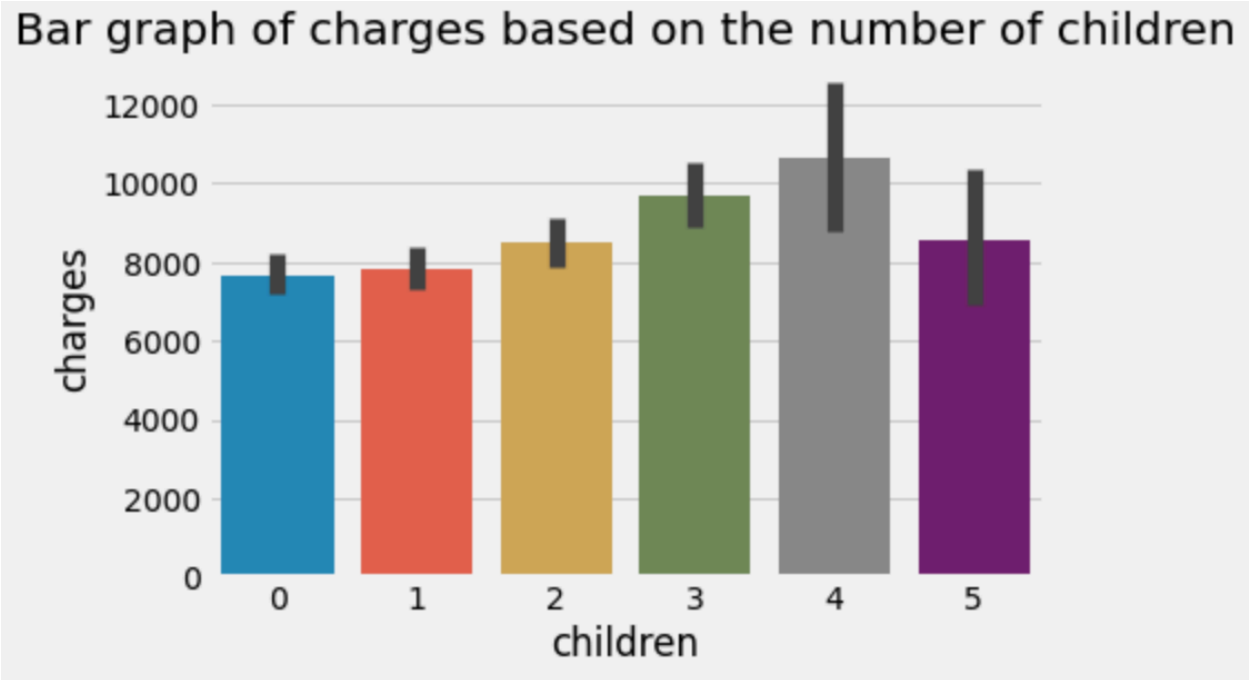


After a closer look at the scatter plot above we can observe that there is no relationship between BMI and charges.

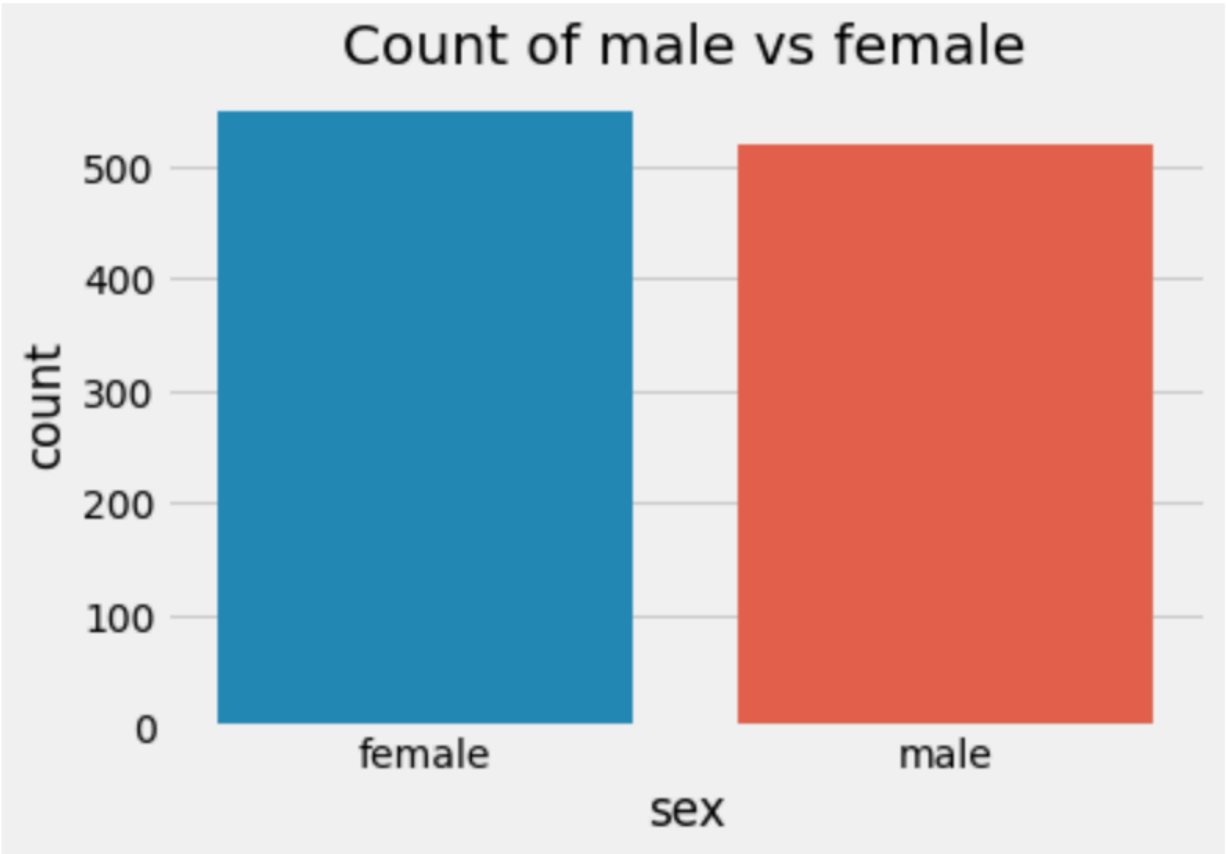
Below is a pie chart to show the distribution of the smokers against the non-smokers and it can be seen that the majority of the subjects are non-smokers to smokers by 93% - 7%.



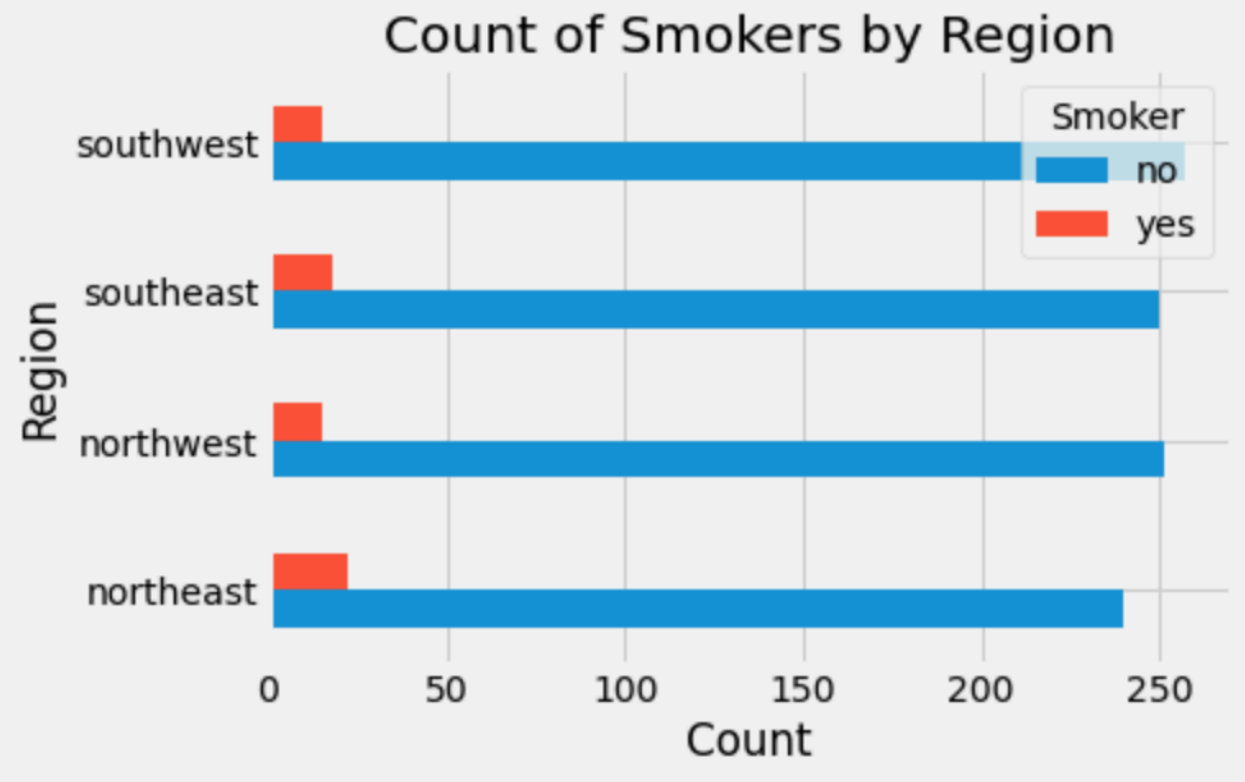
From the Bar Graph below we observe that the charges of the Insurance cost seems to increase until after 4 children and then the charges begin to drop down.



We can also see that the distribution of male and female is evenly distributed.



The horizontal bar graphs show the number of smokers and non-smokers based on their regions. We can observe that north-east has the most smokers followed by south-east and north-west and south-west have equal number of smokers.



4. Methods

The methods implemented in this problem was to encode the categorical data into integers. For the region we chose “get dummies function” from pandas and for the smoker and sex we chose the label encoder function from Sci-Kit learn even though it is usually used for response variables but we can use it for these two variables because they have two distinct values in the variables. I split the dataset into 80% training and 20% testing

Hypothesis Testing

My first hypothesis testing is that does the BMI of males differ significantly from that of females?

So, I stated the Null hypothesis that they are equal and the alternative states that they are not equal.

$H0$: μm = μf

HA: μm != μf

```
Number of males: 519
Variance in BMI of males: 34.028748736082804
Number of females: 549
Variance in BMI of females: 34.473085264979204
```

We cannot reject the null hypothesis because the two BMIs are equal and when we did the two way t-test we had a p-value of 0.57 which is greater than 0.05 so BMI is not statistically significant.

My first hypothesis testing is that does the smokers differ significantly from that of non-smokers?

So, I stated the Null hypothesis that they are equal and the alternative states that they are not equal.

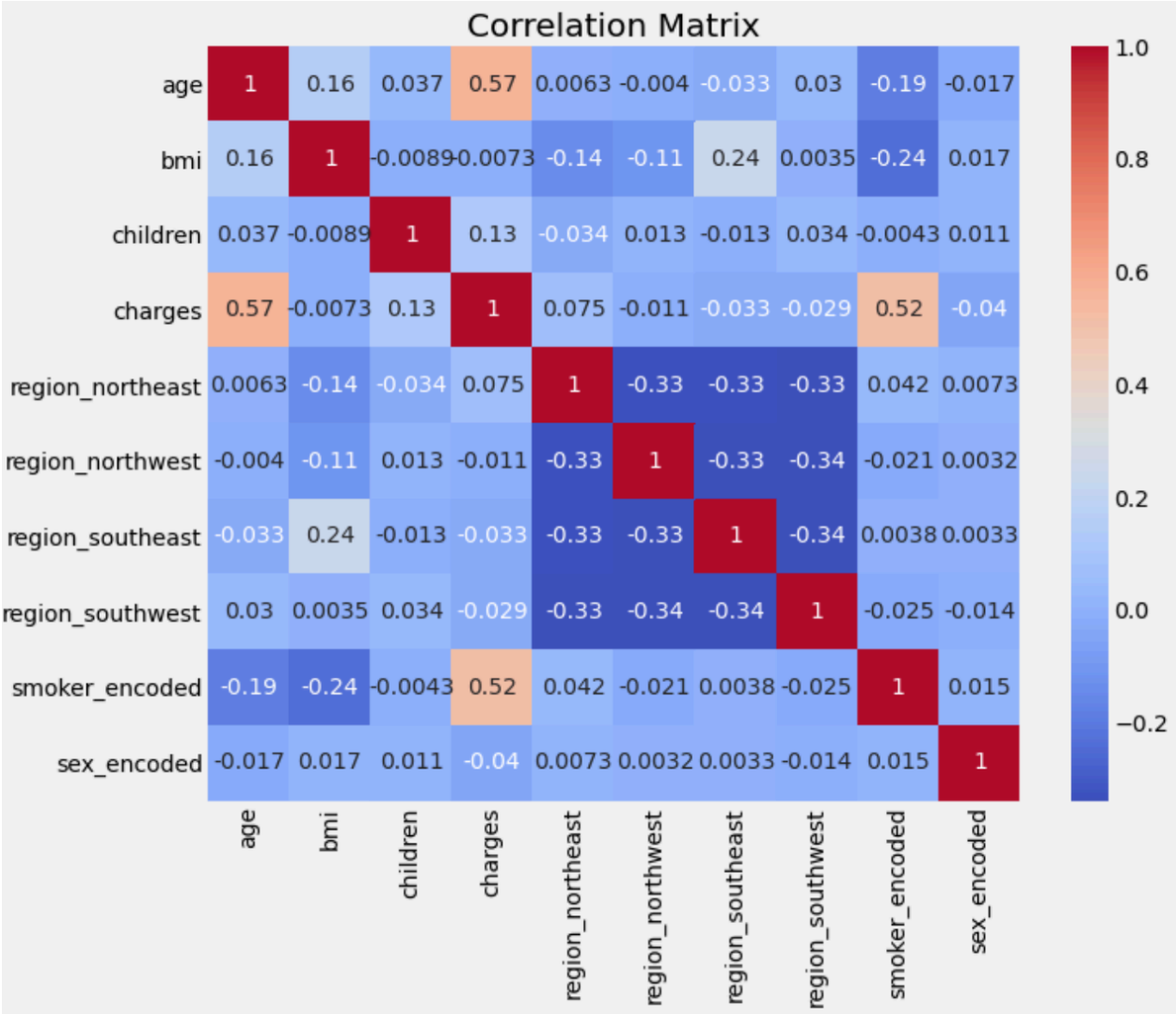
$H0$: μs = μn

HA: μs != μn

```
Number of smokers: 70
Variance in charges of smokers: 3624725.643744107
Number of non - smokers: 998
Variance in charges of non - smokers: 18807962.63411162
```

We reject the null hypothesis because the two variances are not equal and the p-value is 1.37e-71 which is less than 0.05 thus smoking is statistically significant.

We set-up a threshold of 0.3 and a correlation matrix to determine the relevant features based on correlation and we can see that age and smoker are the most important features.

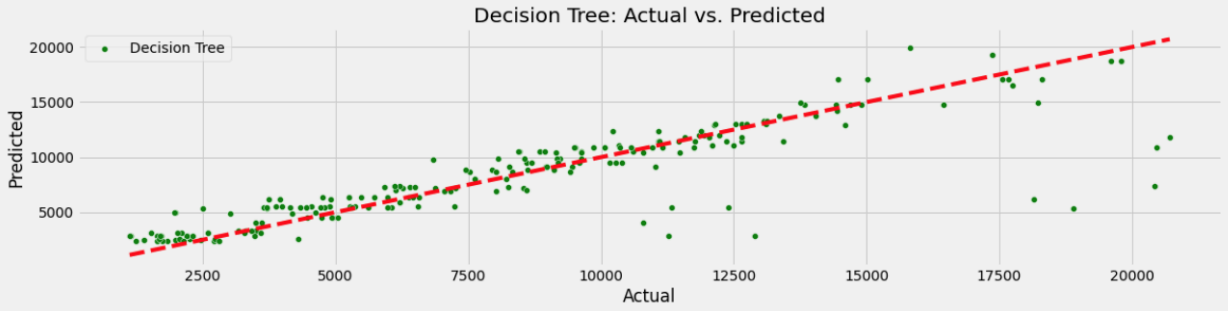



```
Relevant features based on correlation:
['age', 'smoker_encoded']
```

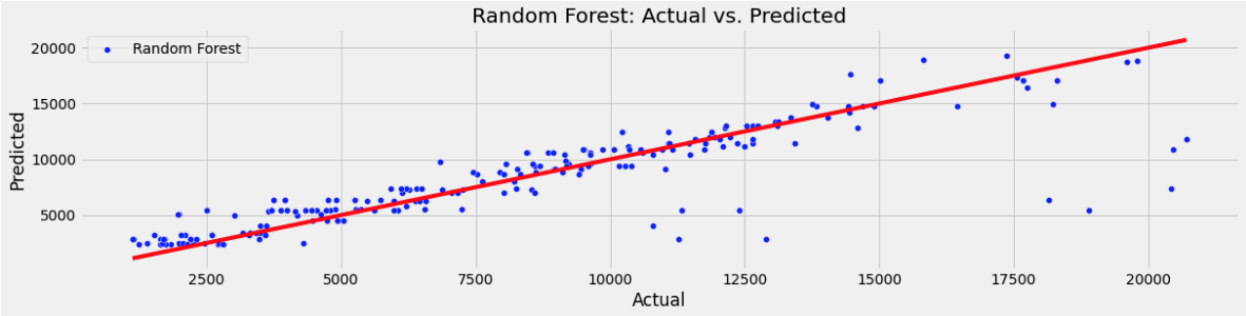
Only age and smoking turned out to be relevant feature based on correlation analysis

5. Data Analyses and Main Results

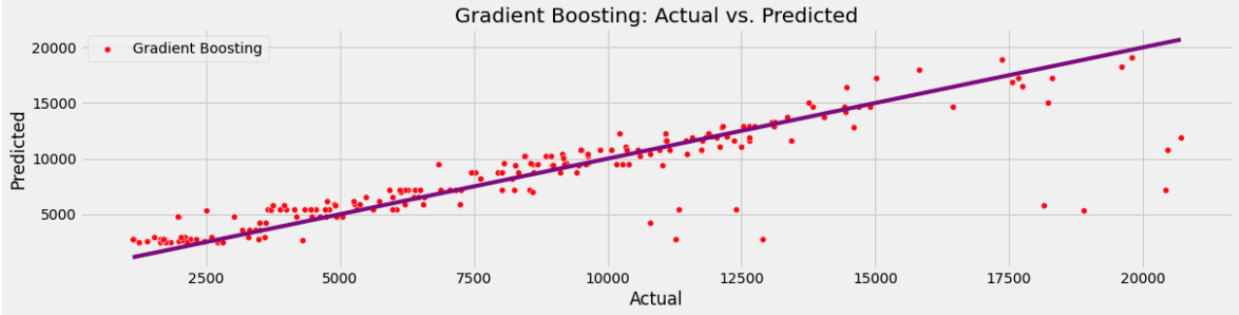
Below is the tree regression actual versus predicted values.



Below is the forest regression actual versus predicted values.



Below is the gradient boost regression actual versus predicted values.



From all these models the predicted and the actual values are not too far off from each other and this can be observed as the regression lines lie on most of the test data in all the models and it looks like the three and the boosting models are performing a little better than the random forest model but this telling is not definite since we are solely looking with the eyes but with metrics we can definitely see the performance of each individual model.

MSE, MAE and R2 Score

Model	MSE	MAE	R2 Score
Decision Tree	5.72863e+06	1275.02	0.746832
Random Forest	5.73423e+06	1264.96	0.746585
Gradient Boosting	5.62008e+06	1234.97	0.75163

From the MSE and MAE and R2 scores are very similar, not much of a difference. The mean squared error measures the squared distance between what my predictor predicted and the actual value and by looking at our MSE and MAE values, the gradient boosting is my best model because it gives the lowest values. my coefficients of determination, R2 scores of 0.7468, 0.7466 and 0.7516 respectively and they are very good also because they show how well the data fit my models. And it fits all the models very well but again showing the Gradient Boosting being my best.

Important Features in Each Model

Decision Tree:

Feature	Importance
age	0.6283321204573394
smoker_encoded	0.3716678795426605

Random Forest:

Feature	Importance
age	0.6119671538381884
smoker_encoded	0.3880328461618115

Gradient Boosting:

Feature	Importance
age	0.6199252235871212
smoker_encoded	0.3800747764128788

In all the models above we can see that age variable is more important than the individuals smoker status.

Feeding my Models new data

When I fed my models new data to see how well it performed they did extremely well as they were able to the insurance charges of a 19 year old smoker with a about 98% accuracy.

Model	Predicted Value
Decision Tree	16483
Random Forest	16484
Gradient Boosting	16486.6

6. Conclusions and Research Directions

Based on my findings and results, I can confidently state that the most important variables in determining the insurance charges for an individual is their age and smoking habits.
Also my Gradient Boosting model is the best performing model which is expected because it is an ensemble method which combines “weak” (high bias) models in an ensemble to collectively lower the bias of the individual models. Overall, the analysis of the insurance charges dataset highlights the potential of machine learning techniques in predicting insurance costs. Future research efforts should aim to enhance model interpretability, although in these models we were able to deduce the most important features but having a full understanding of how the models arrive at their decisions could help avoid overfitting and is crucial for stakeholders, such as insurance companies and policyholders. Explaining the decisions made by complex models like Gradient Boosting could be a focus for future research. Techniques like SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) could be explored to provide insights into individual predictions.
Optimize predictive performance through gridsearch and hyperparameter tuning, and address ethical considerations to develop robust and reliable insurance pricing models. And Finally, to possibly collect more variables, more data and to include time series and longitudinal analysis to track changes in insurance charges over time could provide valuable insights into trends and

patterns. This could involve analyzing historical data spanning multiple years or decades to understand how factors influencing insurance charges have evolved.

Appendix:

https://www.kaggle.com/code/subhakarks/medical-insurance-cost-analysis-and-prediction

https://www.healthdatamanagement.com/articles/how-insurers-can-use-data-analytics-to-make-Etter-decisions

https://www.experian.com/blogs/ask-experian/factors-that-affect-health-insurance-premium-costs/

https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance/data

https://datauab.github.io/medical_insurance/

https://jupyter.uri.edu/user/stouray/lab/workspaces/auto-0/tree/DSP562