Health Insurance

Tuyen Huynh

What is the study aiming to investigate and why is it important?

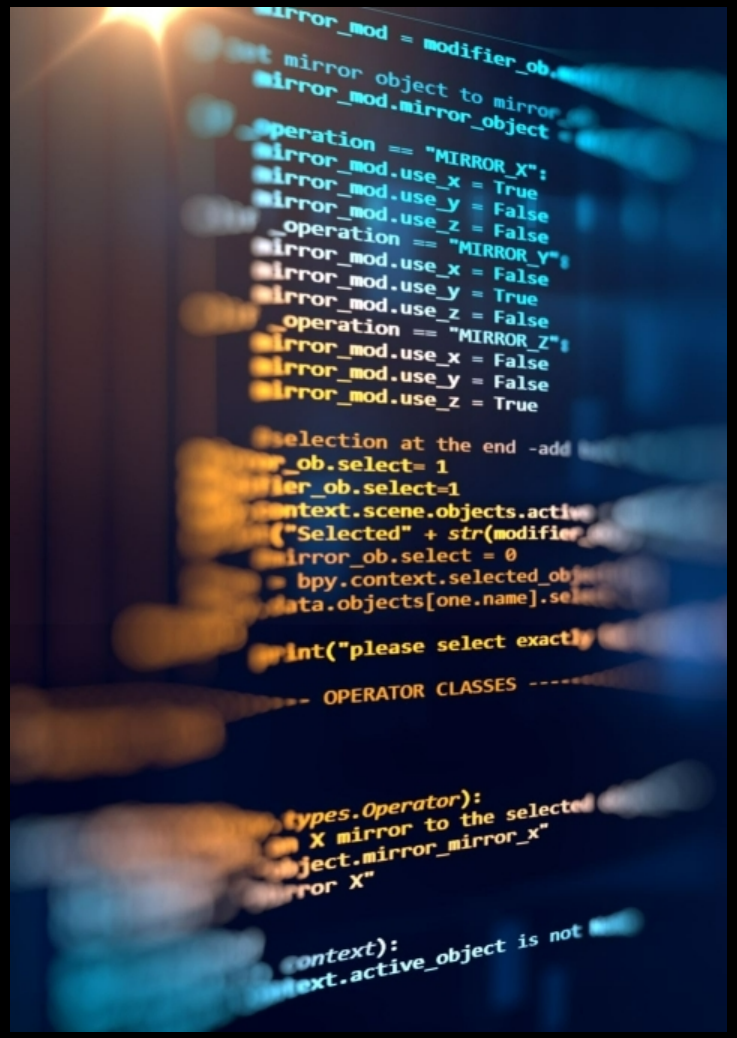
- In this study, we are trying to see what factors play a role in the cost of insurance. It is important to understand why someone might be paying more for insurance than others.

Which dataset is used to investigate?

- The dataset that I am using is an insurance dataset.

What methodologies are used?

- Some methodologies used are the linear regression and the decision trees.



Where was this data obtained?

- The data that I gather was taken online from a website called Kaggle.

- Kaggle is an online community platform that allows users to collaborate with each other.

Data Summary

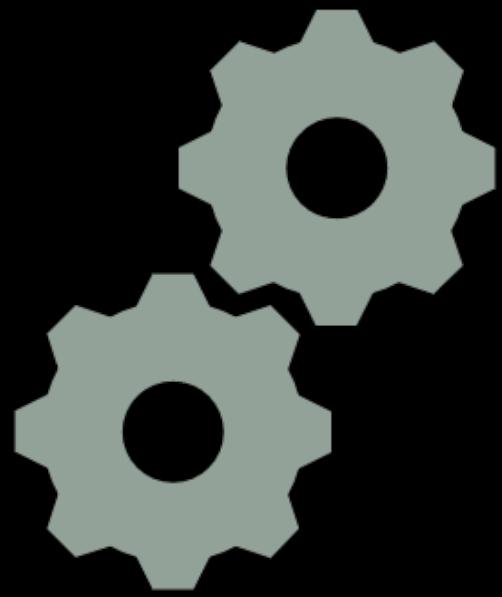
1338 data points

7 variables:

- Age: how old the person?
- Sex: female or male?
- Bmi: body mass index
- Children: number of dependents
- Smoker: smokes or no?
- Region: residential area in US
- Charges: individual cost billed by health insurance

Methodology

- Linear model was utilized to create a linear equation that describes the correlation of the between independent and dependent variables.
- Independent variables: age, sex, bmi, children, smoker, region
- Dependent variable: charges



Methodology

- Decision trees were also utilized to help evaluate options
- Creates a flowchart of the variables to better understand and interpret the data

Regression Results

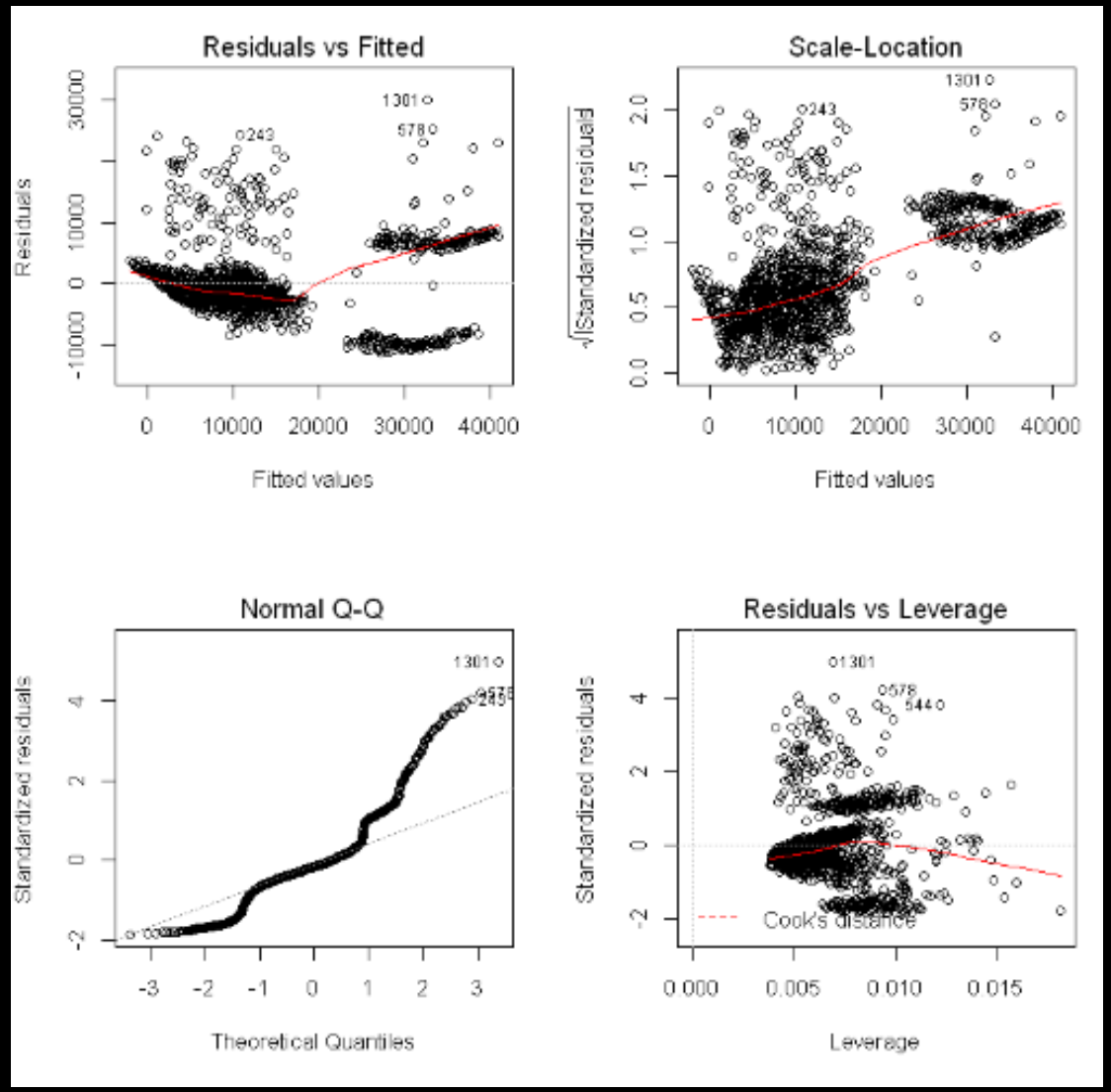
- p-value is less than 0.05 (95% confidence level)
- R-squared is 0.7508
- Data is skewed right (normal q-q)
- Possible outliers: 243, 578, 1301
- Nonlinear (scale-location)

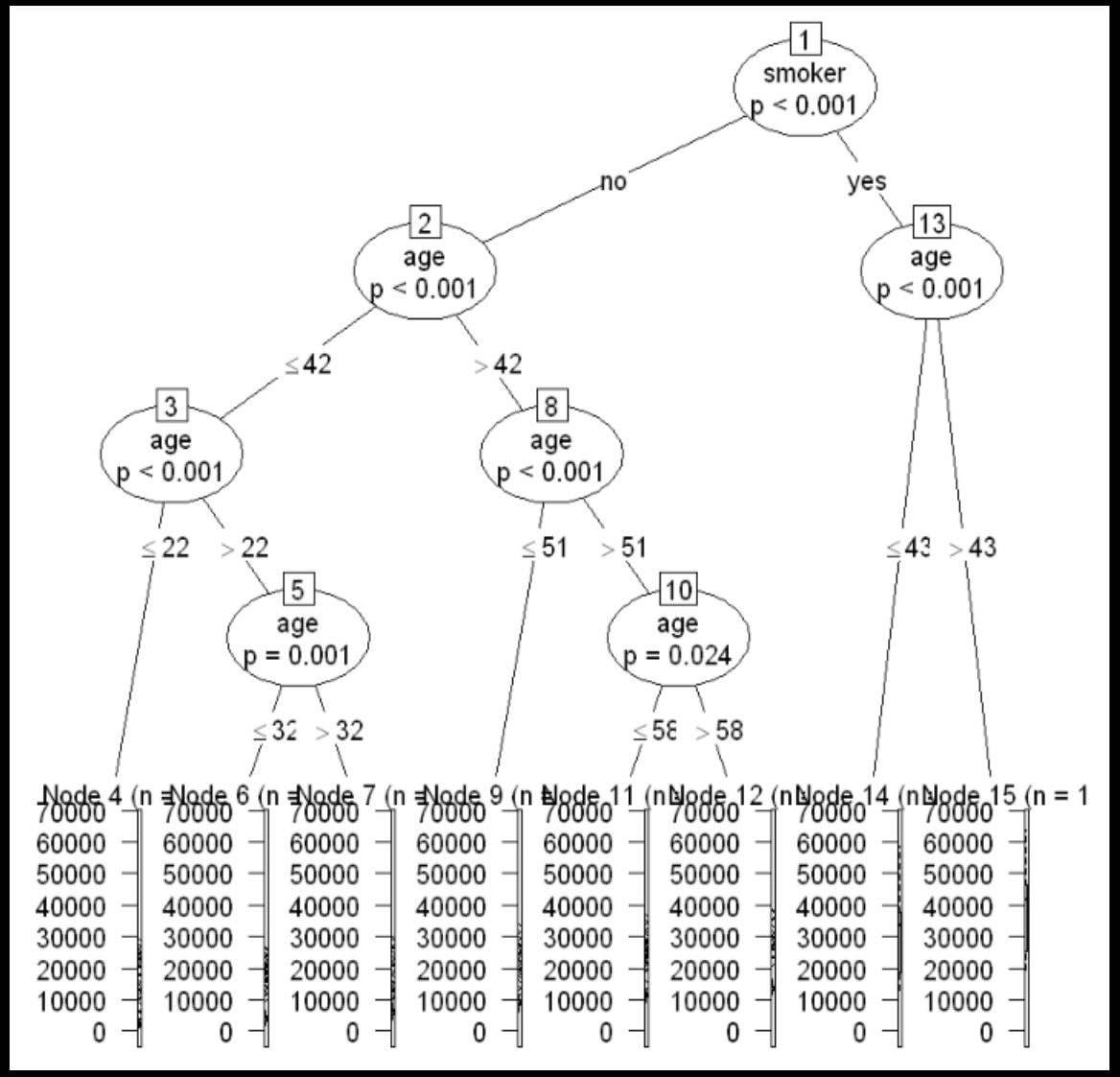
```
Call:
lm(formula = charges ~ age + bmi + children + region + sex +
    smoker)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5       137.8   3.451 0.000577 ***
regionnorthwest   -353.0       476.3  -0.741 0.458769
regionsoutheast  -1035.0       478.7  -2.162 0.030782 *
regionsouthwest   -960.0       477.9  -2.009 0.044765 *
sexmale           -131.3       332.9  -0.394 0.693348
smokeryes        23848.5       413.1  57.723  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```





Decision Trees Result

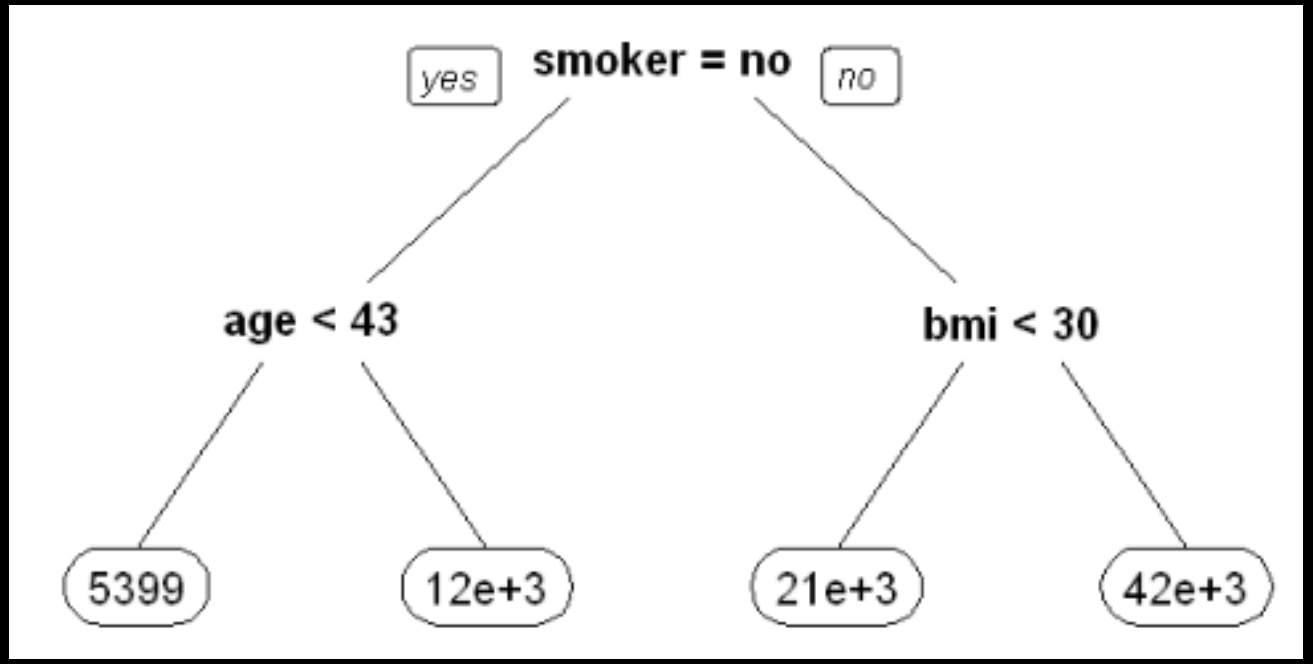
Smokers pay more

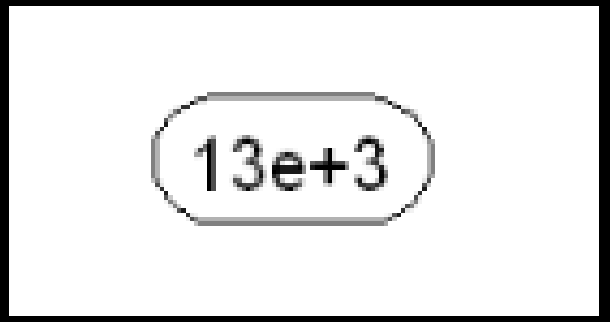
Age affects the pay

Younger people pay less compared to older people

More decision trees results

- Same thing
- Bmi also affects the cost
- Decision trees weren't created for variables: region, sex, and children (overfitting)





Summary

Liner Regression was used for each independent variable; R^2 for bmi: 0.03934, R^2 for region: 0.006634, R^2 for sex: 0.003282, R^2 for children: 0.004624, R^2 for smoker: 0.6198

This means that smoker has the most correlation alone. However, when paired together they can correlate with the dependent variable strongly.

If you smoke, you are expected to pay more for insurance.

If you are younger than 43, you will pay less than those that are older than 43.

If your bmi is lower than 30, you will also pay less than those that have a higher bmi.

The cost of medical insurance is very dependent on the person's overall health.