DATA VISUALIZATION PROJECT

By Rajarshi Chowdhuri , A19026

About Dataset

Dataset Name: Medical Cost Personal Dataset

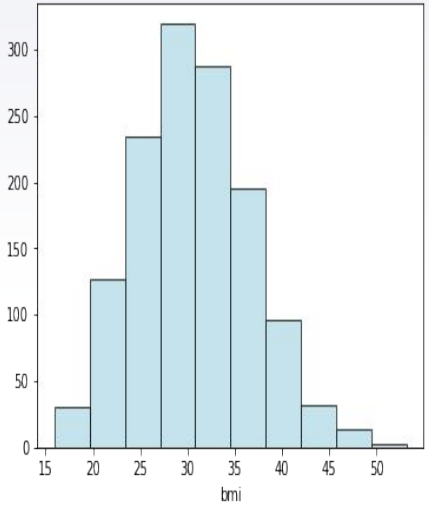
Source: Machine Learning with R by Brett Lantz

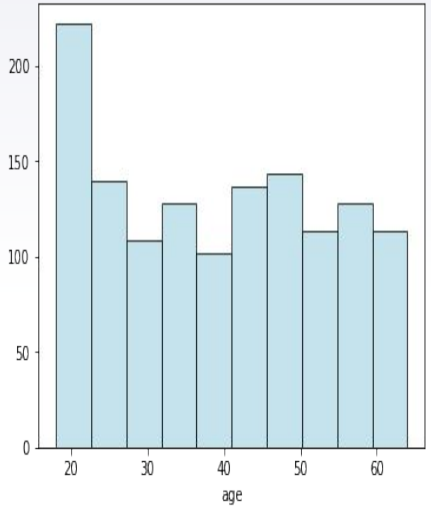
Content:

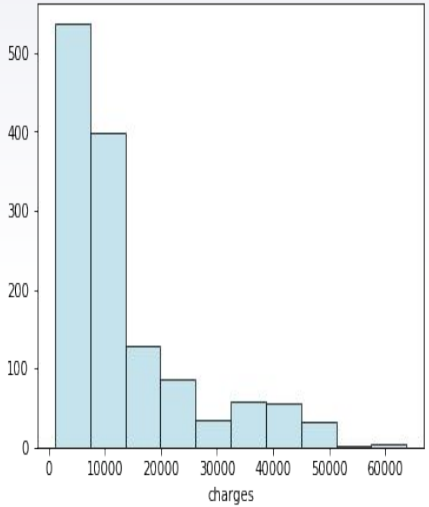
Columns –

- Age: Age of primary beneficiary
- Sex: Insurance contractor gender, female, male
- Bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- Children: Number of children covered by health insurance / Number of dependents
- Smoker: Smoking
- Region: The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- Charges: Individual medical costs billed by health insurance

Histogram



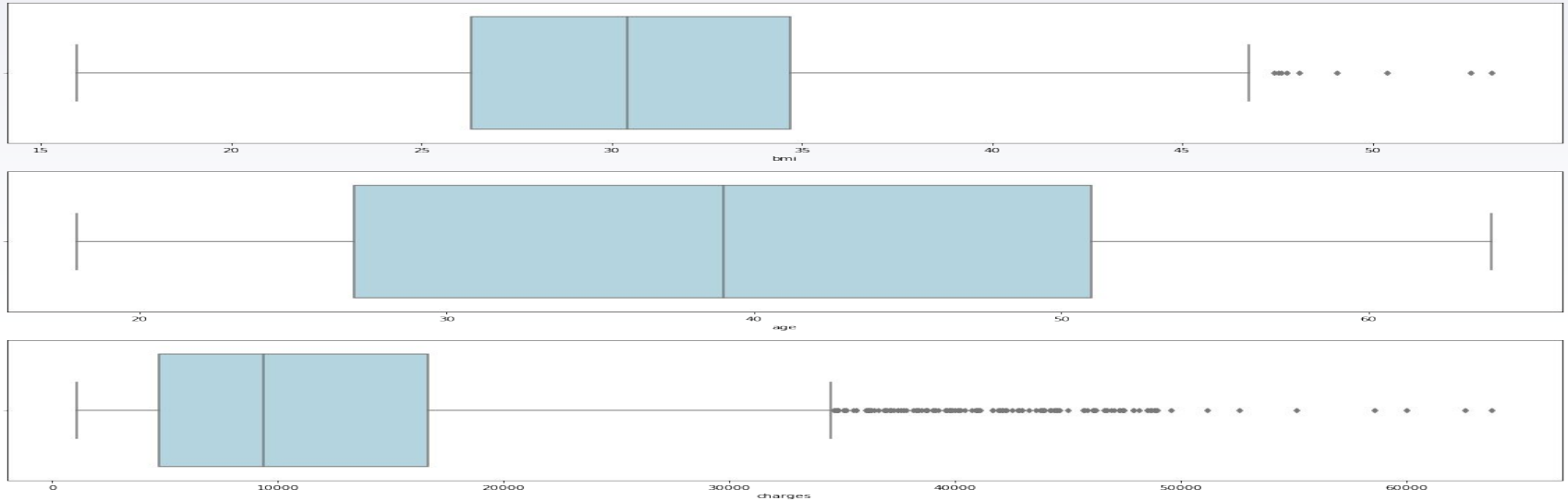




Inference:

- Bmi looks quite normally distributed
- Age seems be distributed quite uniformly
- Charges are highly skewed to the right

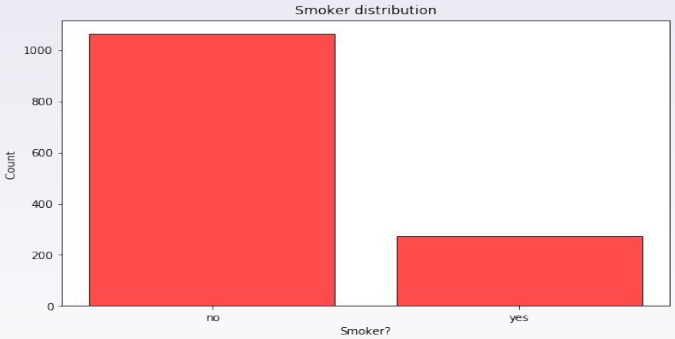
Boxplot

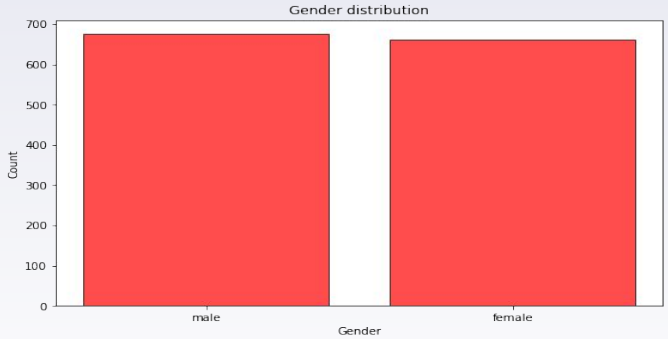


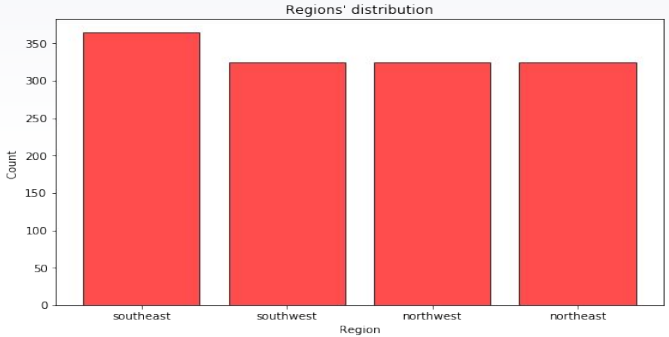
Inference:

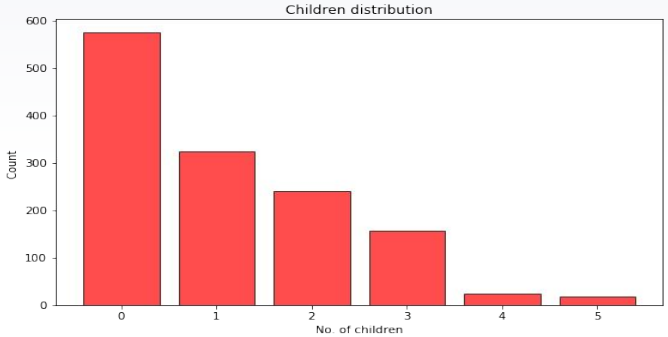
- Bmi has a few extreme values
- Age seems to be quite well distributed and has got no outliers
- Charges as it is highly skewed, there are quiet a lot of extreme values

Bar graph





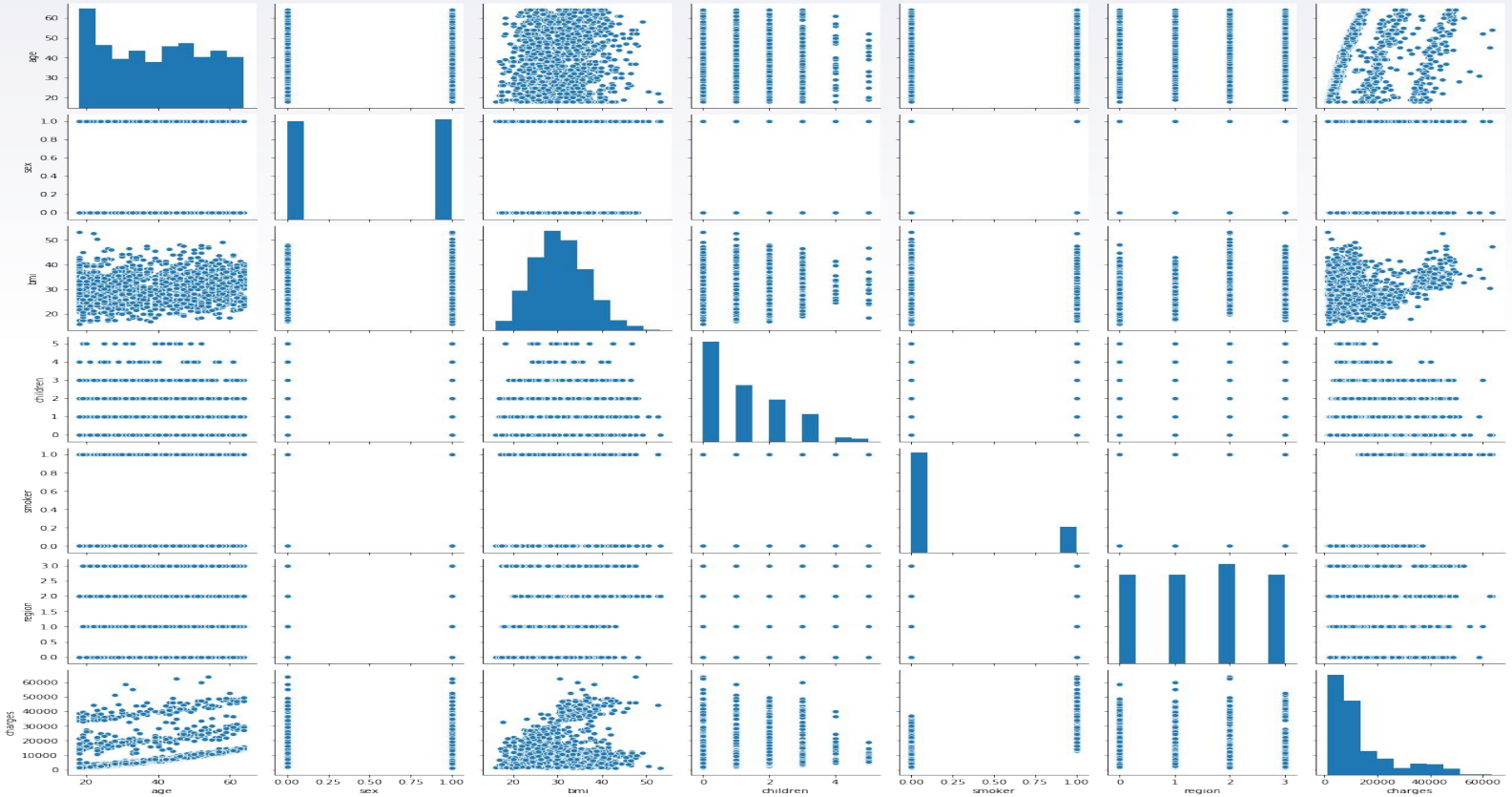




Inference:

- There are a lot more non-smokers than there are smokers in the data
- Instances are distributed evenly across all regions
- Gender is also distributed evenly
- In most instances have less than 2 children and very few have 4 or 5 children

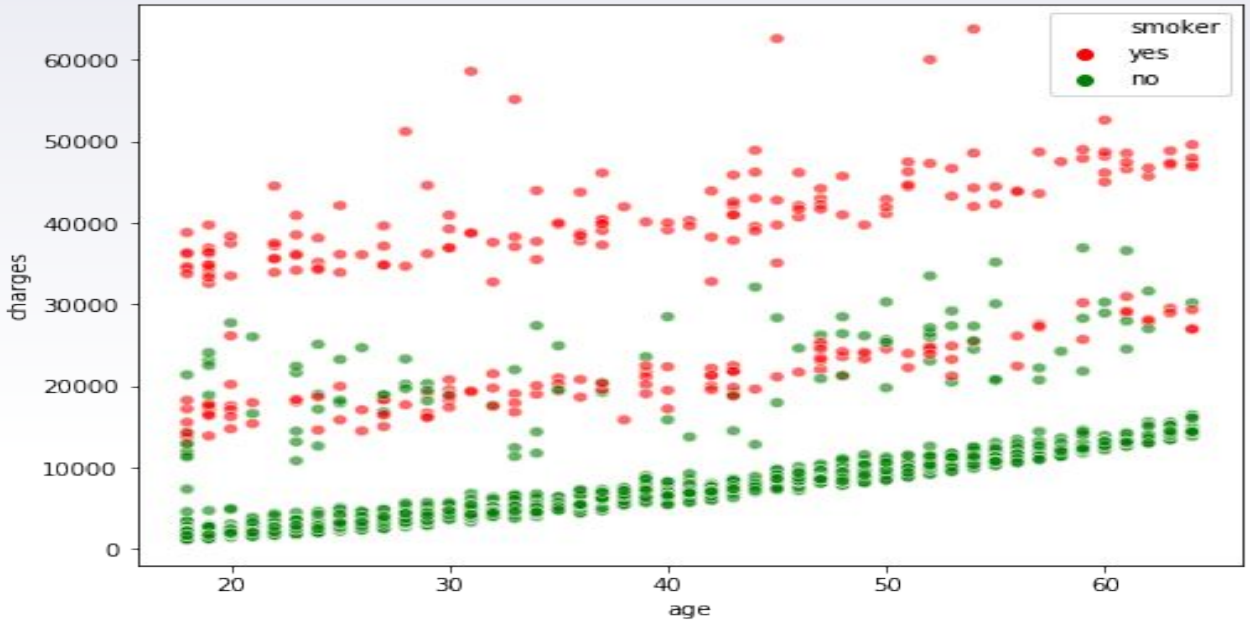
Pairplot



Inference:

- The only obvious correlation of 'charges' is with 'smoker'
- Looks like smokers claimed more money than non-smokers
- There's an interesting pattern between 'age' and 'charges. Could be because for the same ailment, older people are charged more than the younger ones

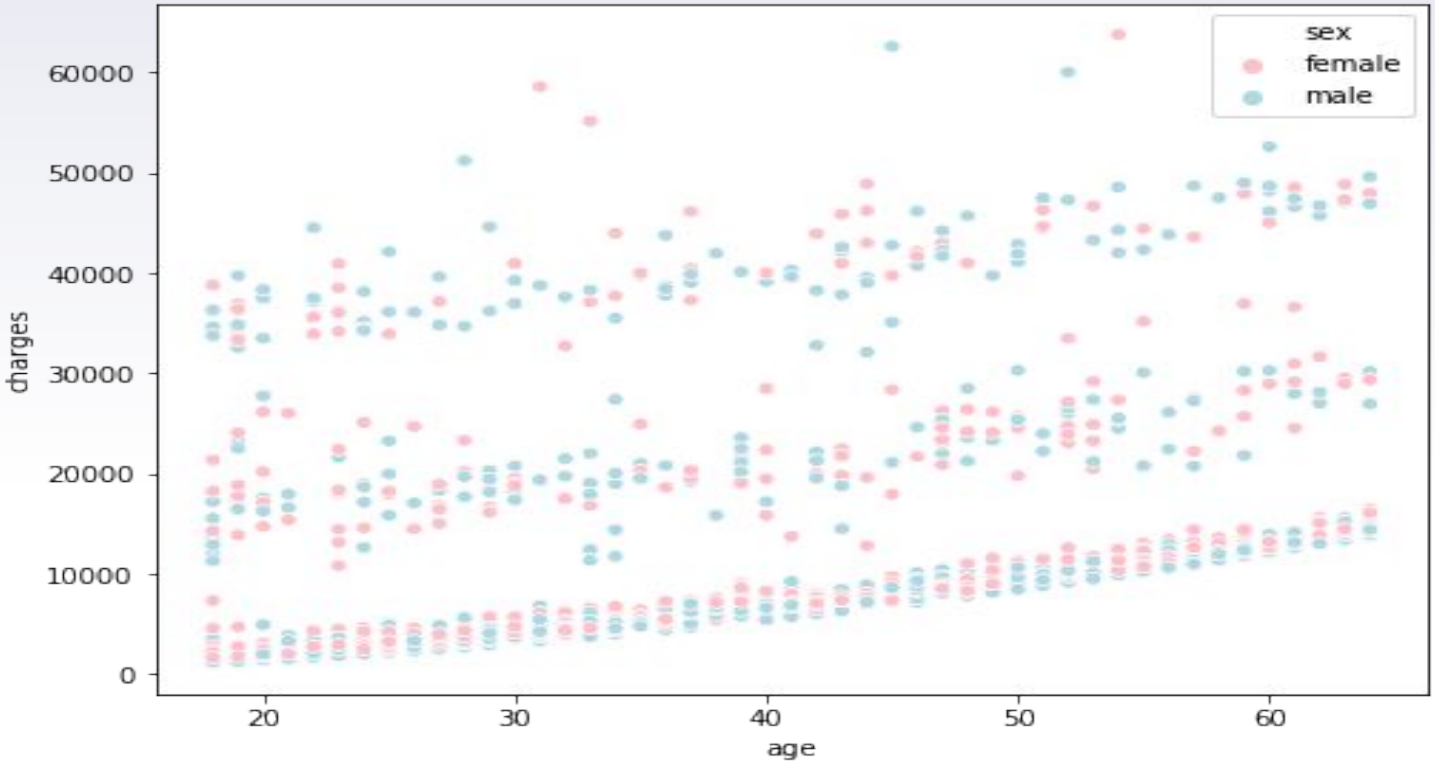
Scatterplot



Inference:

- Visually the difference between charges of smokers and charges of non-smokers is apparent.
- Charges for smokers are much higher than that of non smokers and they constitute maximum of the upper half.

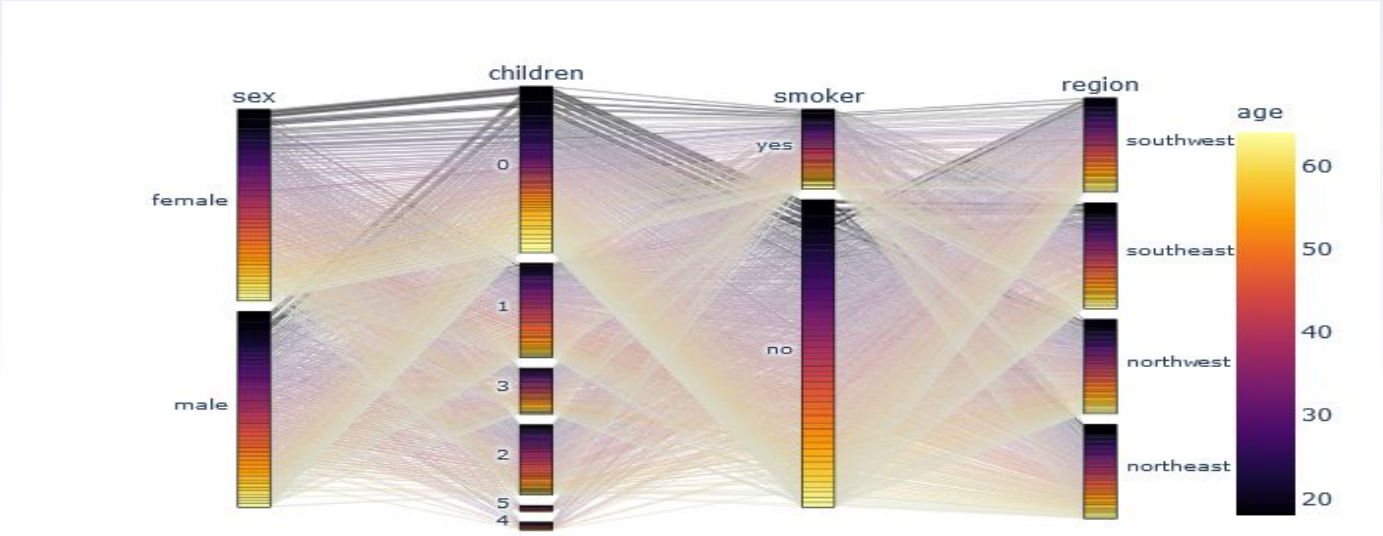
Scatterplot



Inference:

Visually, there is no apparent relation between gender and medical charges, apparently both the genders suffered from diseases in almost equal proportions and treatment charges.

Parallel graph



Inference:

- More number of people hailing from south east has no children
- No of smokers seems fairly high in people of twenties
- Male smokers are higher in number
- Southwest region has least number of smokers

Thank you..

The complete notebook can viewed by clicking this link

Library's used for visualizations in the notebook:

- Matplotlib
- Seaborn
- Plotly