SREE SANKARACHARYA COMPUTER CENTRE
Since 1995
SREE SANKARACHARYA COMPUTER CENTRE PVT. LTD.

Toc H International Academy

PRAGYAN AI
Grow with Gyan

Prediction-of-Health-Insurance-Charges-using-regression

ASWIN.G.KUMAR

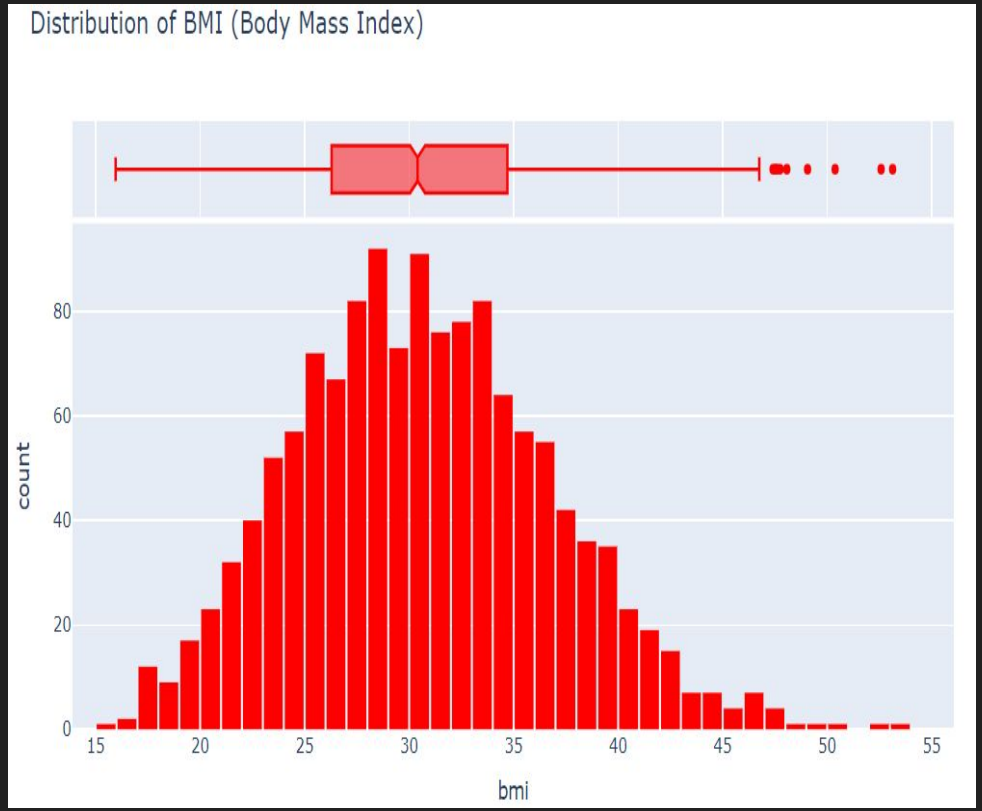
Introduction

This is a health insurance charges prediction model that uses a linear regression algorithm to predict the health insurance charges of a person based on the given data. The dataset used for this tutorial is available on Kaggle and GitHub.

- *This dataset contains detailed information about insurance customers, including their age, sex, body mass index (BMI), number of children, smoking status and region. Having access to such valuable insights allows analysts to get a better view into customer behaviour and the factors that contribute to their insurance charges. By understanding the patterns in this data set we can gain useful insight into how age,gender and lifestyle choices can affect a person's insurance premiums.*
- *This could be of great value when setting up an insurance plan or marketing campaigns that target certain demographics. Furthermore, this dataset provides us with an opportunity to explore deeper questions such as what are some possible solutions for increasing affordability when it comes to dealing with high charges for certain groups?*

Why the distribution of ages forms a uniform distribution while the distribution of BMIs forms a Gussian distribution

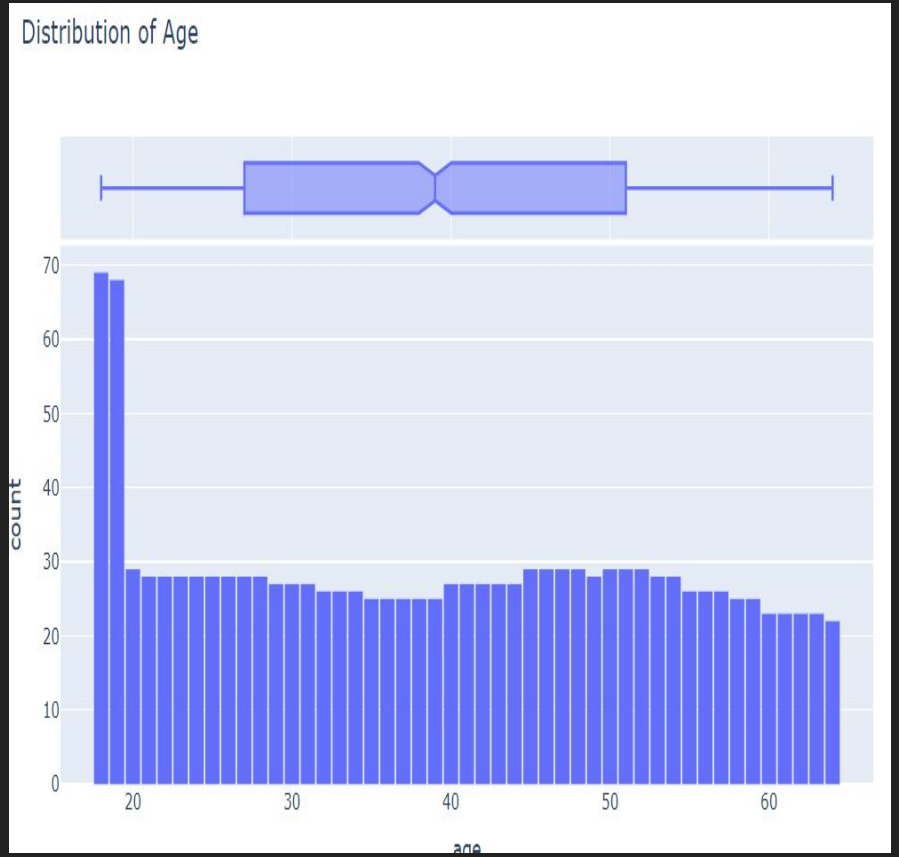
Let's look at the distribution of BMI (Body Mass Index) of customers, using a histogram and box plot.



As there are same number of people in age groups the distribution we get is uniform distribution.But, for BMIs distribution we get gaussian distribution as people who are between 20-3- are considered to be healthy

relatively and less than or more than this range is considered to be health risk categories and these people are more prone to health issues and thus company will have to pay more medical bills of this kind of customers and thus they provide these type of people same health insuarance at higher prices and which in-turn attracts lesser people with out of healthy range BMI because they have to pay more

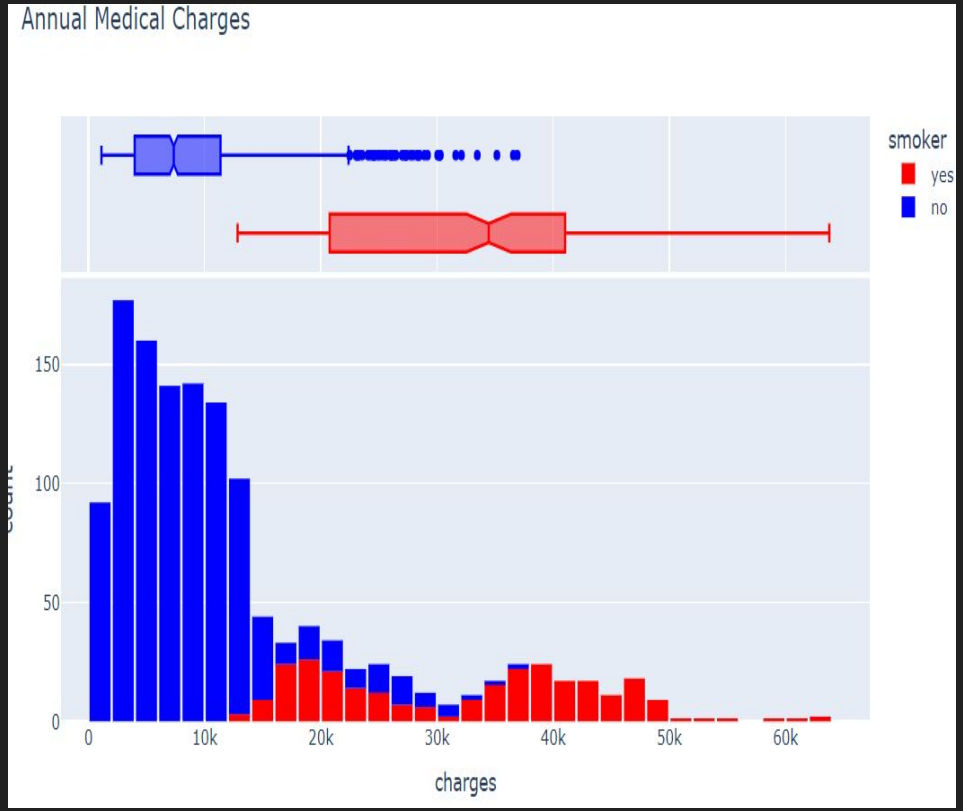
The distribution of ages in the dataset is almost uniform, with 20-30 customers at every age, except for the ages 18 and 19, which seem to have over twice as many customers as other ages.



Why there are over twice as many customers with ages 18 and 19, compared to other ages ?

Insurance agency charges more money for same premium as you get older and here is the same case.people who are younger are less prone to getting sick and thus company has to pay them less for their medical bills.otherwise every age group in US has equivalent population density.

Let's visualize the distribution of "charges" i.e. the annual medical charges for customers. This is the column we're trying to predict. Let's also use the categorical column "smoker" to distinguish the charges for smokers and non-smokers.



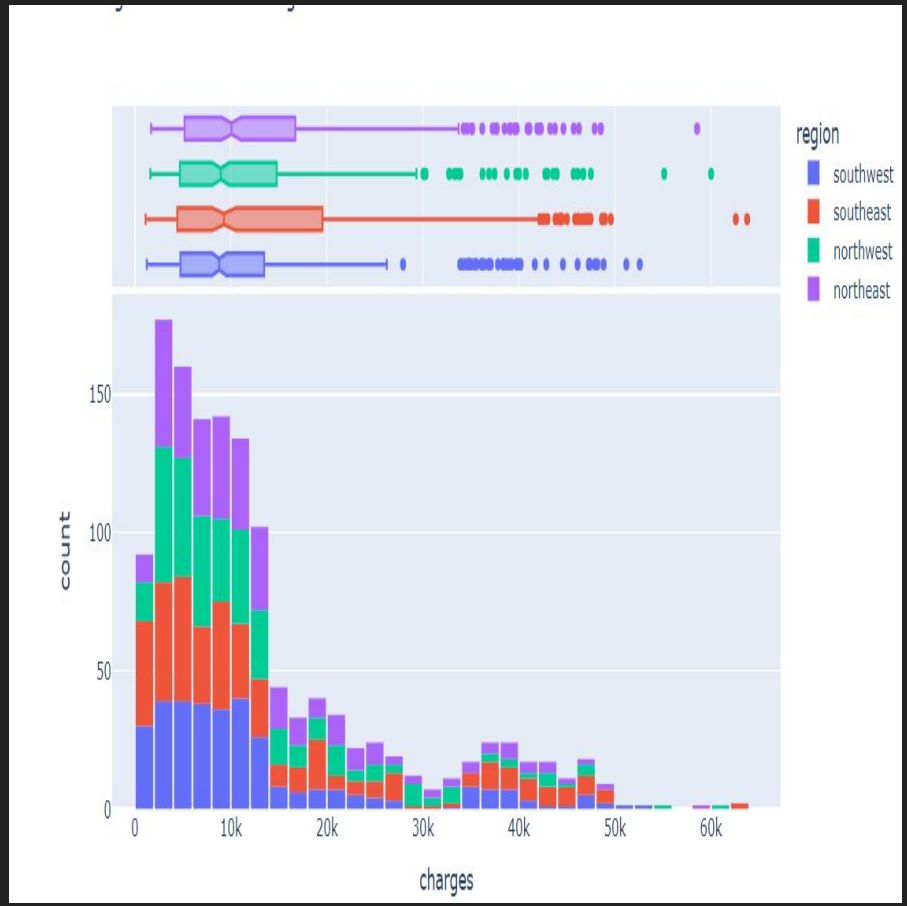
We can make the following observations from the graph:

1. For most customers, the annual medical charges are under \\$10,000. Only a small fraction of customer have higher medical expenses, possibly due to accidents, major illnesses and genetic diseases. The distribution follows a "power law"

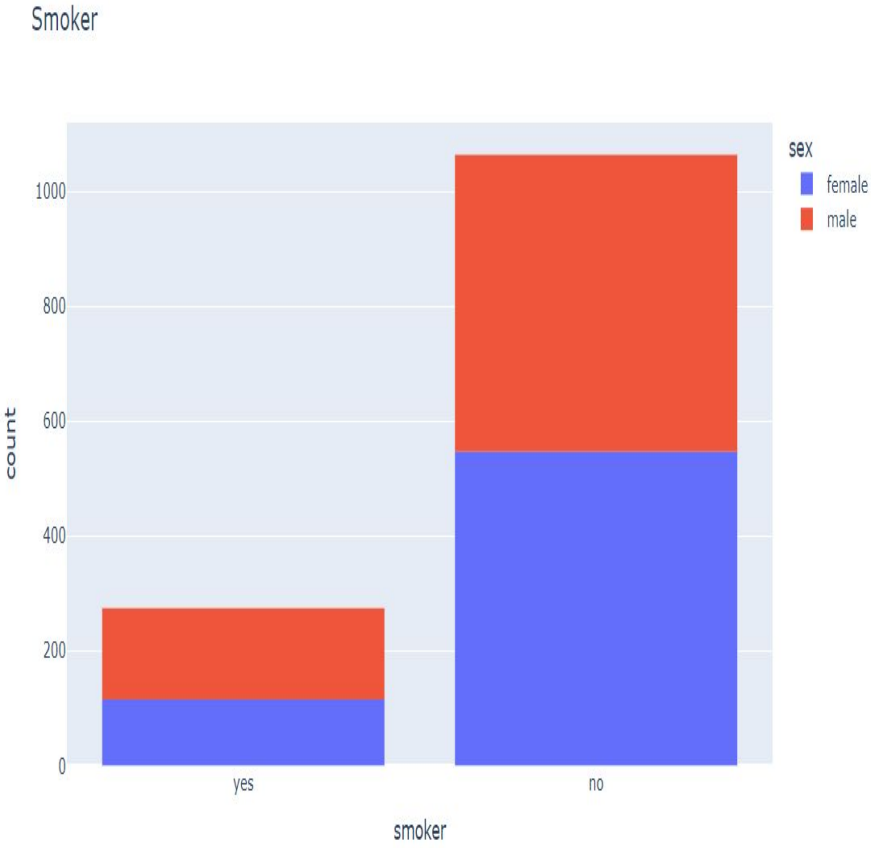
2. There is a significant difference in medical expenses between smokers and non-smokers. While the median for non-smokers is \\$7300, the median for smokers is close to \\$35,000.

How to find the Charges over different regions

Here in the distribution of charges over gender we see that males are substantially charged more because by subconscious behaviour males are exploratory and they are more likely to take risks and that keeps them in danger more than their counterpart.It is also evident that males of U.S are more inclined to get health insurance then female. And, in second distribution we see that south eastern part of U.S is leading in charges but majority of all customers from all parts of US are charged between 0-20k only.

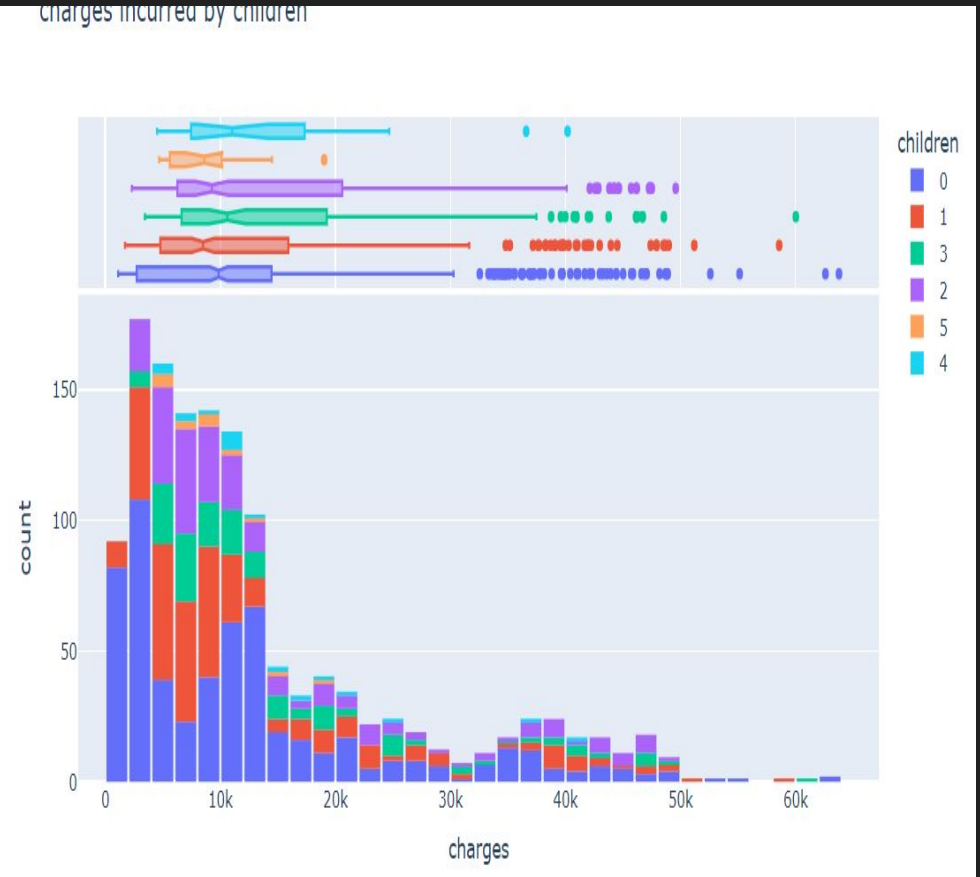


Let's visualize the distribution of the "smoker" column (containing values "yes" and "no") using a histogram.



It appears that 20% of customers have reported that they smoke.We can also see that smoking appears a more common habit among males. This is true for the given dataset only and so we should always verify if these results of analysis also matches the general population which we are going to use our model on otherwise the model will assume that in general population also 20 % people are smokers but in reality it was only 10 % so we would get incorrect predictions.So it is best to check if our primary data analysis matches the results of the general public

How to find the charges incurred by children

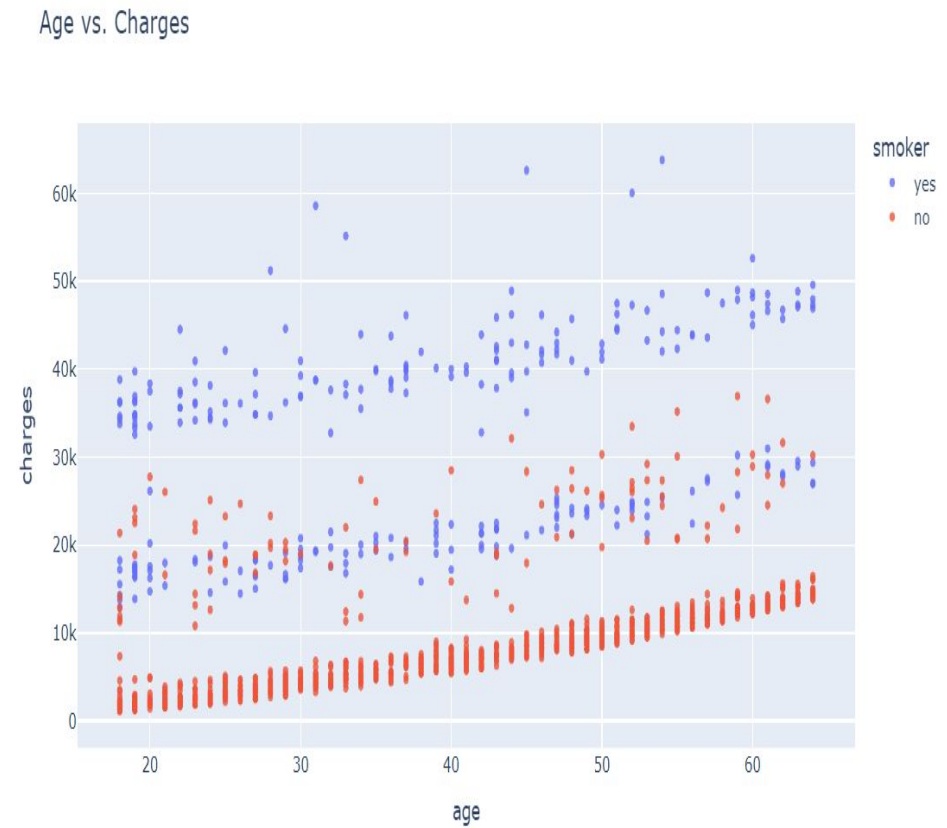


- It seems that majority of our customers have 0 or 1 child and median charges vary between 8.5k to 11k dollars

- We can also conclude that people who have more children are given less priority in terms of pricing discounts

- Having looked at individual columns, we can now visualize the relationship between "charges" (the value we wish to predict) and other columns

Let's visualize the relationship between "age" and "charges" using a scatter plot. Each point in the scatter plot represents one customer. We'll also use values in the "smoker" column to color the points.



*** The general trend seems to be that medical charges increase with age, as we might expect. However, there is significant variation at every age, and it's clear that age alone cannot be used to accurately determine medical charges.**

*** We can see three "clusters" of points, each of which seems to form a line with an increasing slope:**

1. The first and the largest cluster consists primary of presumably "healthy non-smokers" who have relatively low medical charges compared to others

2. The second cluster contains a mix of smokers and non-smokers. It's possible that these are actually two distinct but overlapping clusters: "non-smokers with medical issues" and "smokers without major medical issues".

3. The final cluster consists exclusively of smokers, presumably smokers with major medical issues that are possibly related to or worsened by smoking.

Conclusion

In conclusion, this project on predicting health insurance charges has highlighted the intricate relationship between various personal and demographic factors and the resulting healthcare costs. By analyzing key variables such as age, gender, medical history, lifestyle choices, and geographic location, we have developed a robust model that not only estimates annual healthcare expenses but also offers insights into the underlying trends affecting these costs.

The findings underscore the importance of a comprehensive understanding of individual risk profiles when it comes to health insurance pricing. As healthcare costs continue to rise, insurers and policymakers can leverage our predictive model to create more equitable and tailored insurance plans. This can lead to better resource allocation, improved patient outcomes, and ultimately, a more sustainable healthcare system.

Future work could explore the integration of additional factors such as socioeconomic status and environmental influences to refine our predictions further. By continuously updating our model with real-world data, we can enhance its accuracy and applicability, paving the way for informed decision-making in health insurance and healthcare delivery.

Overall, this project not only serves as a valuable tool for stakeholders in the health insurance industry but also contributes to a greater understanding of how personal choices and circumstances impact healthcare costs.