Setting

Business Context

As the world continues to recover from a global pandemic brought on by the COVID-19 virus, and as people continue to face various health complications and issues, there is no doubt that health insurance has turned into a necessity for much of the population. In fact, over 92% of the U.S population has health insurance. Since insurance has clearly become such a hot commodity in today's world, large companies have entered the market, and have come up with different ways to gain their customers' business. In this project, we have chosen to use cluster analysis to assign customers with similar characteristics to their own clusters, which in turn allows insurance companies to charge different groups of customers differing rates, depending on their characteristics.

Problem

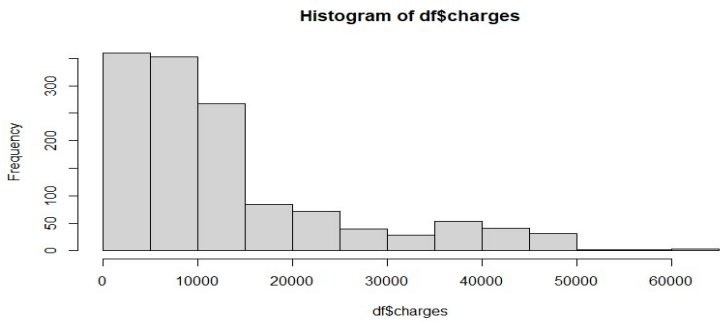
Through this project, we are hoping to take the dataset that we have, use cluster analysis (unsupervised learning) to allocate the customers into 'k' clusters. The goal of this analysis is to cluster the similar patients and set the premium depending on their characteristics.

Data Description

Our data contains 7 columns, and below are the details of it:

Variable	Data Type	Description
age	int	age of beneficiary
sex	chr	Policy holder's gender: male/female
bmi	num	body mass index
children	int	number of children covered in the insurance
smoker	chr	Whether the insured smokes tobacco regularly: yes/no
region	chr	four regions: northwest, northeast, southwest, southeast
charges	num	individual medical costs billed by insurance

The data includes 1,338 examples of beneficiaries currently enrolled in the insurance plan, with features indicating characteristics of the patient as well as the total medical expenses charged to the plan for the calendar year. The source of data set is from the Kaggle, in which data was created from the US census bureau. There are no missing values in the data.
From charges, we can observe that a large majority of individuals in our data have yearly medical expenses between zero and $15,000.



We have 676 male and 662 female beneficiaries in the data. And beneficiaries are almost similarly spread across the regions. (For code, please refer to #Beneficiaries proportion. There are few more graphs in R code describing the data.)

To perform cluster analysis, a correlation test for the variables, and other tests, we first had to make sure that all the variables were converted to numeric values. This meant that we had to convert the values in the "sex", "smoker", and "region" variables to numeric values.

For sex, female = 1 and male = 0
For smoker, yes = 1 and no = 0
For region, northeast = 1, northwest =2, southeast = 3, southwest = 4

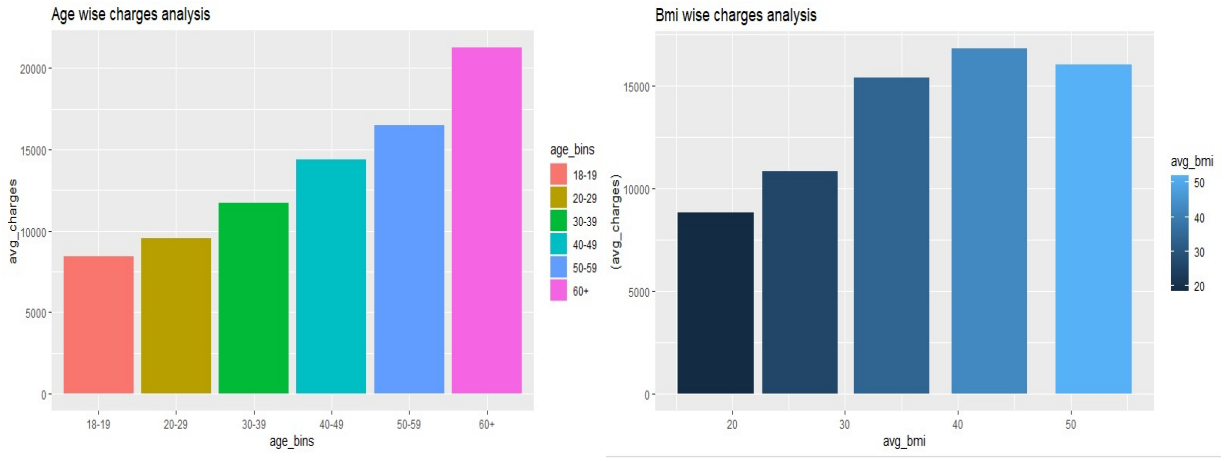
Analysis and discussion

After converting the variables, we ran a correlation test. (For complete result, please refer to code #Correlation between variables)

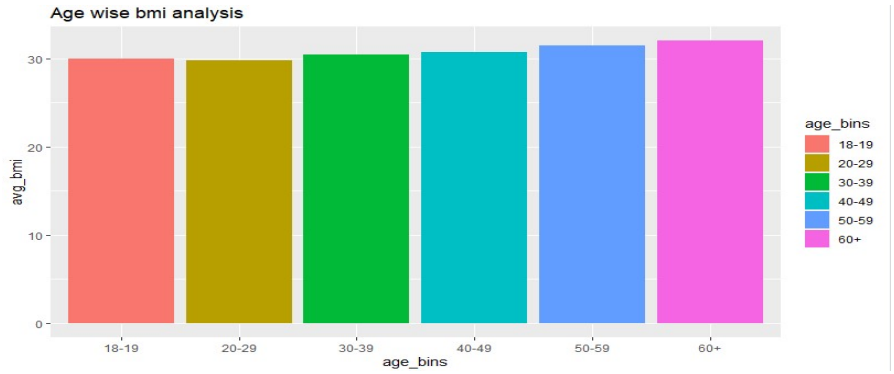
We can see a strong positive correlation between smoker and charges, with which we can infer that smoking can increase medical expenses due to lung diseases.

There is also a positive correlation between age and charges & BMI and charges, which means as a person gets older, the medical charges for the beneficiary increase which is justified as health complications increase with an increase in age. A higher BMI corresponds to obesity which can lead to

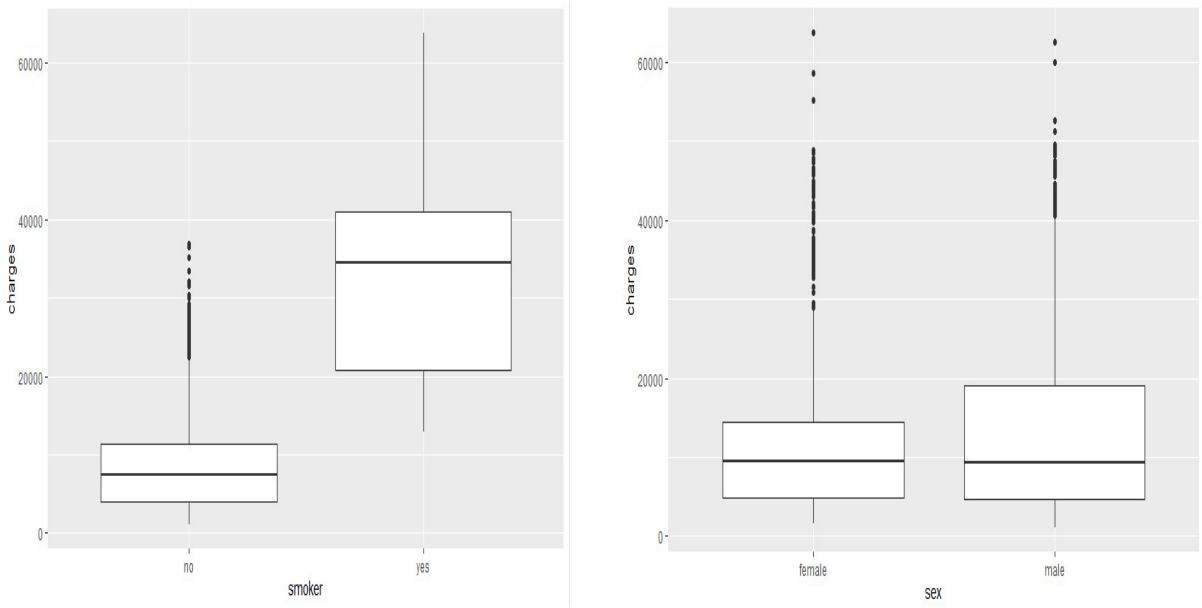
health and medical issues which in turn increases the medical expenses and leads to a higher insurance premium.

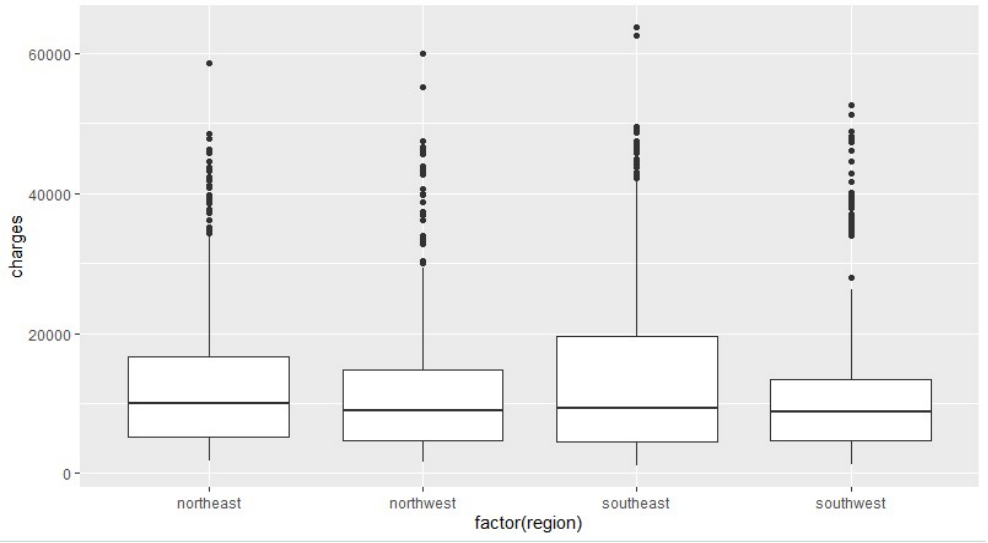


From the plot, we can observe a moderate correlation between age and bmi, as age increases, bmi also increases.



To look out for more clear relationships between variables and charges, below are few box plots.





The average expenses are similar for male and female. But charges are higher for smokers than non-smokers.

According to the regional analysis, the average charges are almost similar across all the regions, although the highest is from the southwest region.

To perform clustering, the data has been normalized. The optimum number of clusters is two which was obtained using the Silhouette score. K-Means clustering is performed, and the resulting two clusters have sizes of 1064 and 274 units.

The resulting cluster stability is 0.3 which is low and there is correlation between few variables. So, we perform Principal Component Analysis to create new independent variables from the data and improve the cluster stability.

We have used Scree plot (plot of the proportion of variance by each PC) to check for the elbow to determine the optimal number of PCs for clustering. From the plot, we can observe the elbow at the second PC, considering it we use PC1 and PC2 for the next process. The weights of variables in PC1 and PC2 are provided in the code. (Plots are available in R code)

Clustering is performed again with the new PCs, resulting in two clusters of size 1055 and 283. Cluster stability has increased to 1.

Adding cluster labels back to the data and visualizing it we found the following results

Cluster 1 (size = 1055) belongs to beneficiaries who have lower charges than beneficiaries from cluster2 (size = 283). And most of the cluster 2 people have smoking habits.

We can also observe that as age is increasing charges are increasing in both the clusters. Both the clusters have people of all ages. (More visualizations are available in R code)

We have also performed Hierarchical clustering using Euclidean distance and Ward's method. Though the order of cluster is different, results are similar.

From the results, we can fix the premium higher for cluster 2 and lower premium for cluster 1. If the company can gather additional information from its customers, then having additional independent variables would lead to more efficient classification of the clusters.

The biggest contributors to an increase in the mean charges happen to be Age, BMI, and Smoker Status in the following ways:

1. The lower the age, the lower the mean charges among the users, not considering other factors like BMI and Smoker status. The mean charges increase linearly with age, with the increase between the age groups being around 10-20.
2. Poor BMI and Smokers appear to be paying a higher mean charge even if their age falls in a lower range. A higher BMI may affect the mean charges by approximately 40% - 80% over the mean charges the other group pays.
3. A user with a good BMI but slightly older in age may end up paying more than a younger user with a poor BMI or is a smoker or both.

Conclusion and Solution

1. Cluster 1 in our study includes more Non-smokers, lower BMI, and lower mean age. Typically, the charges this category of beneficiaries pay are much lower than those with the opposite characteristics. The difference in the charges being paid among the two Clusters can be calculated to be approximately 20 - 50%.

   From studies, a base lower premium can be set for Cluster 1. The base premium will increase depending on the variation among the parameters Age and BMI. For example, the difference between a child beneficiary and an adult, with different ages.

2. For Cluster 2 which majorly includes Smokers, beneficiaries of higher mean ages and a relatively poorer BMI, the charges can be observed to be significantly higher than that of Cluster 1. The variation in charges observed can be up to twice as much.

   The Charges in Cluster 2 are at least 40% greater than Cluster 1 and can even go up to more than 100% higher than Cluster 1. For Cluster 2, setting a higher base premium relative to Cluster 1 would make sense as the applicant or the beneficiary if falls into this cluster category, would most definitely be paying a higher charge. Less variation in the charges can be expected due to Smoking being prevalent in Cluster 2.

From our study, we observe that effective categorization of the applicants or the beneficiaries into Clusters like those we incorporated in our study could help the insurance companies charge the premium amounts more accurately considering the various health parameters of the beneficiaries.