MAKE HEALTH INSURANCE GREAT AGAIN

How can we leverage data and make businesses relevant and competative in the market? Can we take advantage of technology and data to solve socio-economic issues?

AUTHORS

Wai Kong Ng - 20221384
Engjulla Hasani - 20221404
Johnas Camillius - 20220723
Pedro Rodrigues - 20220639

01. INTRODUCTION

Health insurance has been a major problem in US. If they charge too much they lose customers and that's opposite if they don't charge enough, they will go out of business.
Hakuna Matata Insurance is our new health insurance company. How can we price our health insurance packages so that we can be competitive in the market? Our solution would be using R to create Regression Model which will predict how much citizens will pay anually for their medical bills based on some generic features such as age, number of children , body mass index, region as well as smoking habits.

02. DATA EXPLORATION

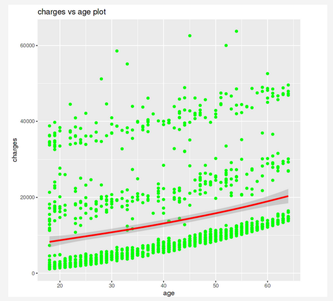
A. We converted no of children to categorical variables.
B. We didn't find any missing values (null values)
C. Our average age was 39, minimum age was 18 where as the oldest person was 64 years old.

Interesting note: We found out that bmi can affect healthcare consumptions price.
Note: bmi = body mass index

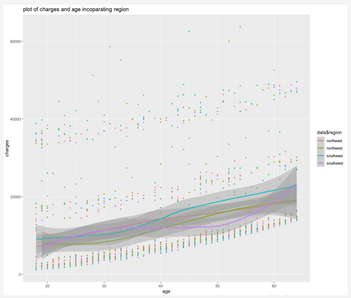
03. ANALYSIS

1. All of the explanatory variables have an effect on charges, according to the analysis. The two biggest interactions in the dataset were bmi and smoker, and there is a strong correlation between log(age) and log (charges).
2. We can see from the data visualization that smoking has a significant impact on costs, followed by age and other factors.
3. The r2 score of the model is 90% on train data and 96% on test data because it can explain 90% of the variability on the train dataset and roughly 96% of the variability on the test data set.
4. The residual analysis shows that mistakes are regularly distributed and do not exhibit any patterns, however the right side exhibits a sizable departure. The variation of errors is nearly constant while using qqplot, however a tiny rise is seen on the right side of the plot.
5. There is around 30% correlation between age and charges and from analysis of plot we are now finding the coorelation between log(charges) and log(age)
6. There is very little impact of sex on charges.
7. No. of children has impact on charges for same bmi but we cannot find any perticular pattern in it.

Graph. 1



Graph. 2



04. MODEL

1. Our model was linear regression.

Note: LR should be applied to data which is normally distributed. We used logarithmic transformation (Graph. 6 & 7) technique to normalize our data. We used qqplot to make sure our data was normally distributed (Graph. 3)
2. We used boxplot to check outliers and we found out that we don't have bad-behaved data in our set.
3. Muliple R-squared was 84% and we wanted to improve our score by removing observations which were creating large errors (residuals). Thanks to plot command which helpedus to identify these residuals.
4. After running again the model we improved it and attain a new score which was 90%
5. We tested our model by 20% of our original data which we split before and have 20% test data as well as 80% training data. So almost 96% of variability of test data is explained by model which is good fit on test data

Note: We wanted to use confidence interval but we realized that it will be helpful if we are using our model for prediction purpose.

04. RESULTS/FINDINGS

From Graph.4 & 5 we can say that there is very high interaction between age and smoking, bmi and smoking so these effects must be taken into consideration while modelling charges.
From prediction vs actual data plot we can notice that the model works excellent for smaller values charges but for larger value of charges the model is not that accurate because of less data for larger claims.

IMPORTANT!
Make healthcare affordable to all.

That is our motto.

06. CONCLUSION

Our health insurance firm will use smoking habits, age, region and bmi variables when pricing for it's products and services.
As age of patients increase their medical bills increase proportionally (Graph.1) We will implement different prices regionally as income is different across multiple regions (Graph. 2)

Further research: Our research answered medical expenses question and helped us to be relevant in the market but we had limitations like how much patients spend annually from public and private service providers? As researchers we has assumptions that private institutions charges more and their patients define more premium consumers.

References

https://blogs.sas.com/content/iml/2017/06/01/choose-seed-random-number.html

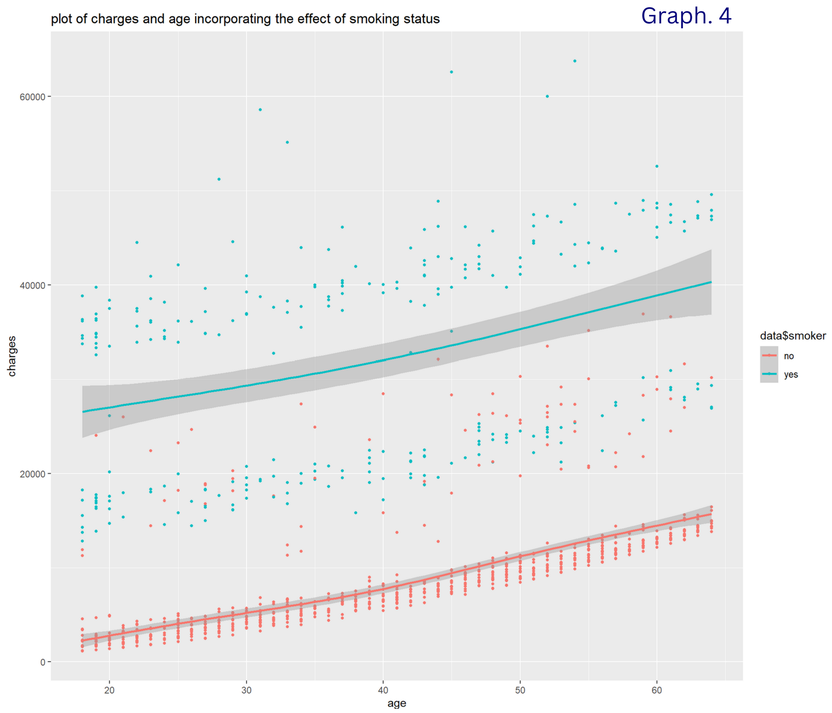
Class notes, theoritical classes and practical.

https://kenbenoit.net/assets/courses/ME104/logmodels2.pdf

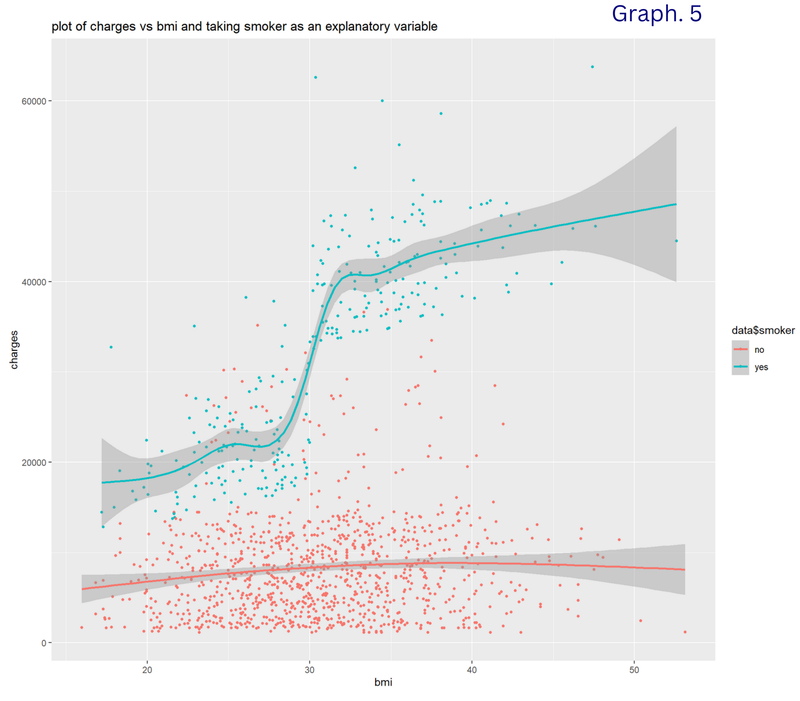
https://pubmed.ncbi.nlm.nih.gov/11323447/

GRAPHS

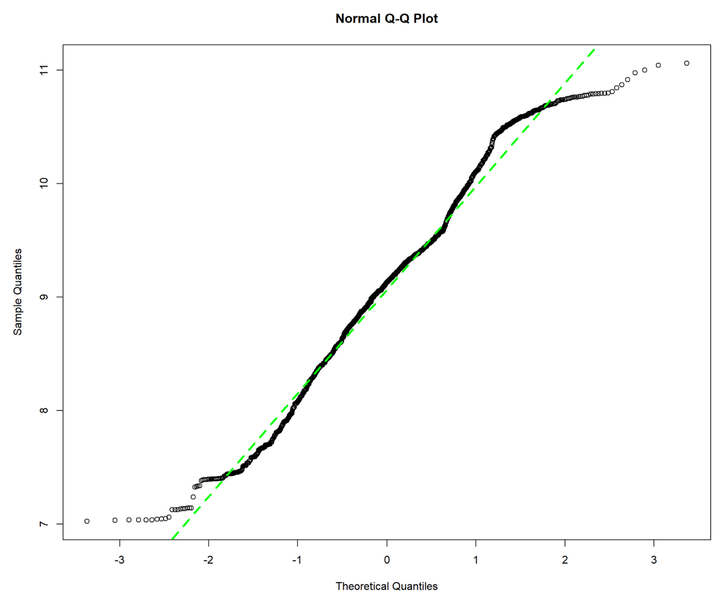
Graph. 4



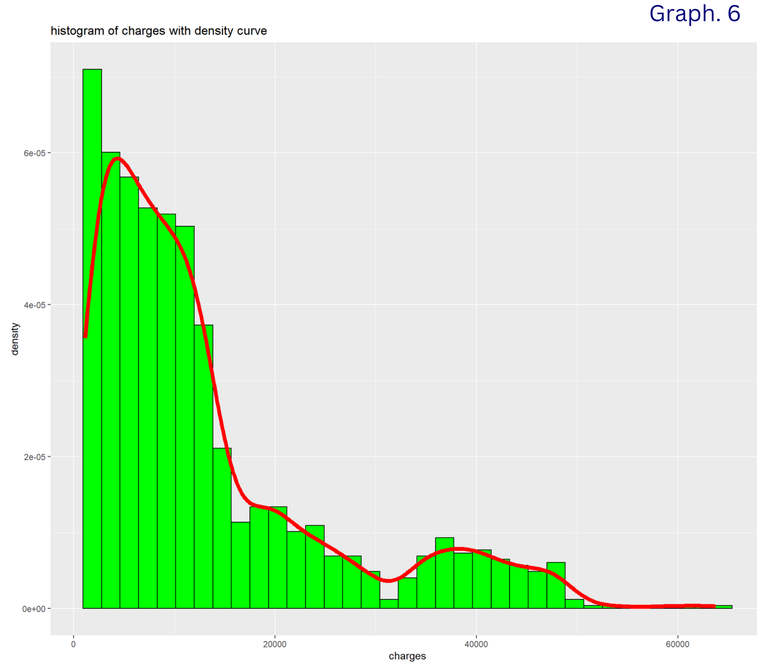
Graph. 5



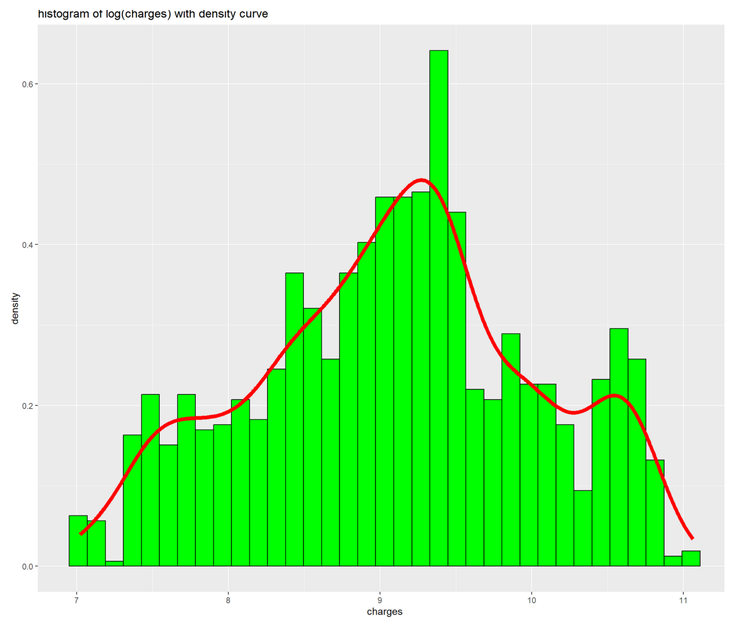
Graph. 3



Graph. 6



Graph. 7



Graph. 2