The Cost of Medical Care

Can general health factors predict insurance billing?

Corey Quackenbush – CS 522 – Data Mining

Introduction

Health coverage in the United States is critically important because unlike many other developed nations, the healthcare system in the US does not provide universal coverage for its citizens [1]. As a result, most Americans purchase health coverage through private third parties and various federal and state programs [2]. Given the current COVID-19 pandemic, it is more important than ever to ensure that people are covered by health insurance so that if they end up needing to go to the hospital, they won't get stuck paying the entire bill. This is even more critical given the fact that US medical charges are higher when compared to other developed nations [3]. This project attempts to uncover the factors that could contribute to higher medical expenses as well as examining their underlying patterns. For example, given a set of attributes, can we predict how high the associated medical bill will be? Also, can we use the data to discover unknow groups that aren't obvious from direct comparisons?

The dataset for this project was created for the book – Machine Learning with R [4] – and is available on GitHub [5]. It was created by sampling demographic statistics from the US Census Bureau. As such it approximates the real-world conditions in the United States and accurately reflects the population at large. The primary benefit of using simulated data is that there is no need to be concerned about personally identifiable information. The general conclusions drawn from this data should be able to be applied to real patient information, and still hold true, if needed. The dataset consists of 1338 entries, where each entry represents a single patient. Since we're interested in understanding general patterns that lead to higher medical costs, the attributes for each entry reflect general information about each patient. In other words, we're not interested in specific diseases – for example – that cause higher medical costs. We're interested in general factors and combinations of factors that lead to higher expenses.

age	
Min.	18.0
1st Qu.	27.0
Median	39.0
Mean	39.2
3rd Qu.	51.0
Max.	64.0

charges	
Min.	1122
1st Qu.	4740
Median	9382
Mean	13270
3rd Qu.	16640
Max.	63770

bmi	
Min.	16.0
1st Qu.	26.3
Median	30.4
Mean	30.7
3rd Qu.	34.7
Max.	53.1

children	
0	574
1	324
2	240
3	157
4	25
5	18

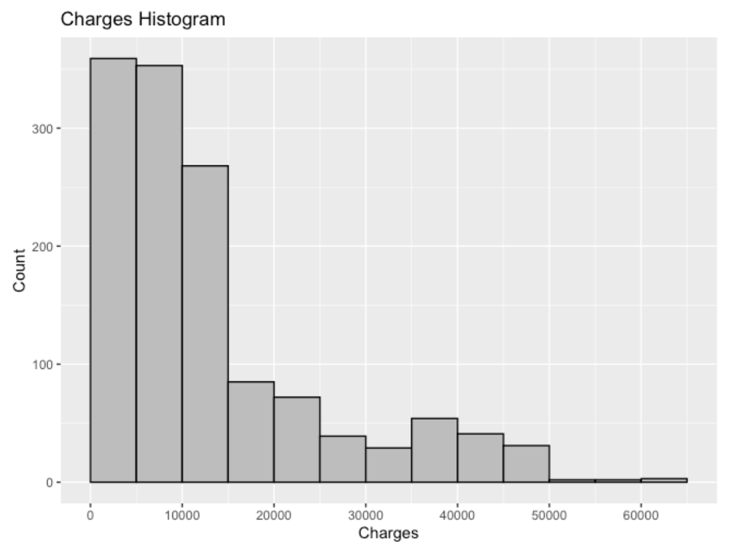
sex	
female	662
male	676

region	
northeast	324
northwest	325
southeast	364
southwest	325

bmi_category	
Underweight	20
Normal	225
Overweight	386
Obese	707

smoker	
yes	274
no	1064

obese	
yes	707
no	631



Methods

The data for this project was analyzed with R [8] through the IDE RStudio [9]. Many of the base packages from R were used to process and transform the data. To gain a better understanding of the data set, summary statistics were generated for each of the attributes. The attribute fields bmi_category and obese were created based on the BMI field that was present in the original data. The CDC guidelines for defining adult overweight and obesity categories were used to determine the cutoffs. Before analysis began, the data was split into a train and test set so that the validity of the models could be assessed. Approximately 90% of the data was used for training and 10% was used for testing. Next, matrix scatter plots were used to determine attribute correlation and general trends were created with the package psych [11]. Note, the Pearson correlation was used by this package to determine the correlation values in the upper right of the plot. Using the information gained from this plot, attributes correlated with medical charges were tested in a variety of regression models. In other words, the lm function within the R base [8] stats package was used to develop linear models to fit the data and to calculate summary statistics about the model. These summary statistics were used to pick the best model. After that, general data exploration was performed by creating plots using the package ggplot [10]. This part of the analysis focused on the attributes that were highly correlated with the cost attribute since that field is what we're interested in predicting. Before attempting to cluster the data, the package, scales [12], was used to perform a linear conversion of attribute values in order to ensure that one attribute set didn't dominate the other. Base R packages [8] were used to perform k-means and hierarchical clustering as well as visualization of the dendrogram. Within hierarchical clustering, Ward's method was used to determine which clusters to merge. The package, DBSCAN [13], was used to create the k-nearest neighbor plot as well as cluster the data points. Elements of the package caret [14] were used to calculate the root-mean-square error and r-squared values to evaluate the regression models on the test data.

Conclusion

In summary, this project is focused on understanding what relationships exist between attributes, finding patterns in the data, and using the patterns to predict outcomes. In this case, the main outcome is being able to predict what price an individual will pay in yearly medical expenses, given a set of general factors associated with them. For example, given a patients age, BMI, number of children, and knowing if they smoke, can be used to predict their annual medical expenses. The model can explain 85% of the variation in medical charges and most predictions are over predicted by only $400 - $1800. This information could be used to help patients budget for future expenses. For example, if the patient has a flexible spending account, they could use this information to increase or decrease their contributions to that account. This could allow them to partition their pre-tax money to pay for their medical expenses. This information could also be used to help convince a patient to become healthier. For example, a doctor could tell their patient how much money they could save by making certain life changes.

Hierarchal clustering using Ward's method was very effective at capturing the obese-smoker group. Once the data set is grouped, new samples can be categorized by either re-running the clustering algorithm or by simply calculating the k-nearest neighbor of the new points to determine which group they belong to. This method is very useful because allows us to derive information about the patient with very little information. For example, if we know the patients BMI and yearly medical expenses, we could say something about how likely they are to be a smoker. Similarly, if we know the patients BMI and whether they were a smoker, we could estimate amount of yearly medical expenses they would pay. Finally, if we knew the patient was a smoker and we knew their yearly medical charges, we could determine if they were obese or assign them to a BMI range.

References

[1] Institute of Medicine. Committee on the Consequences of Uninsurance. 2004. Insuring America's health: principles and recommendations. Washington, DC: National Academies Press.

[2] Millman M, editor. 1993. Access to health care in America. Institute of Medicine, Committee on Monitoring Access to Personal Health Care Services. Washington: National Academies Press.

[3] Ridic G, Gleason S, Ridic O. Comparisons of health care systems in the United States, Germany and Canada. Mater Sociomed. 2012;24(2):112-120. doi:10.5455/msm.2012.24.112-120

[4] Brett Lantz. 2019. Machine Learning with R: Expert techniques for predictive modeling (3rd ed.). Packet Publishing, Birmingham, UK.

[5] Machine Learning with R: Datasets. GitHub. https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv

[6] Tan P, Steinbach M, Karpatne A, Kumar V. 2019. Introduction to Data Mining (2nd ed.). Pearson Education Inc., New York, New York.

[7] RDocumnetation: Correlation, Variance And Covariance. Retrieved from: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor

[8] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

[9] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. http://www.rstudio.com

[10] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

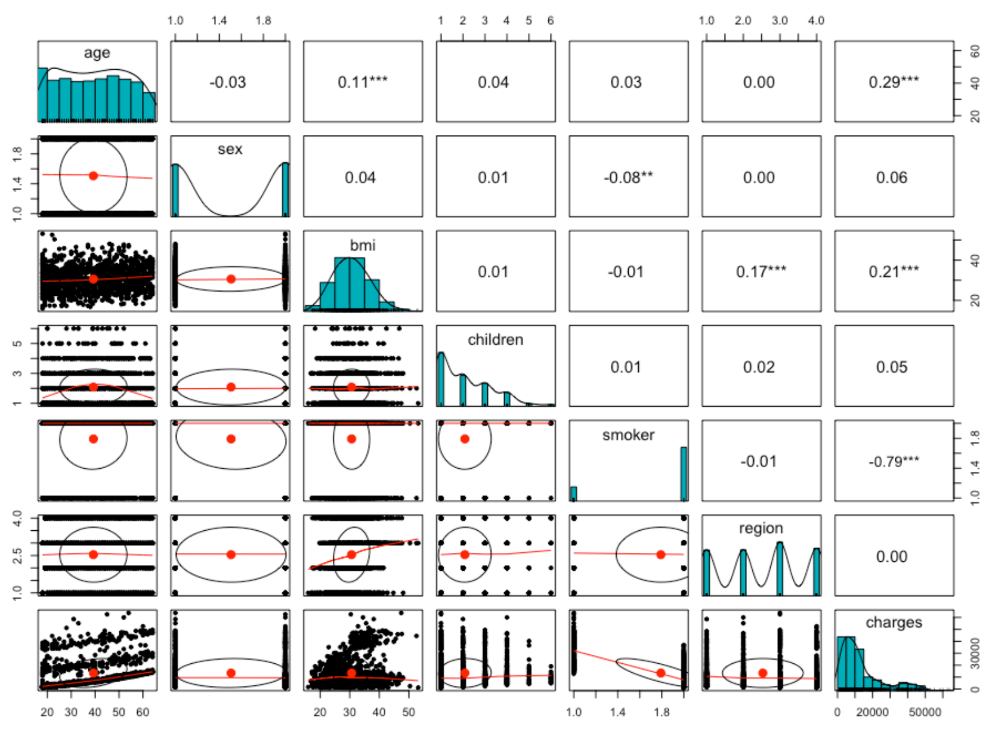
[11] Revelle, W. (2020) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA.

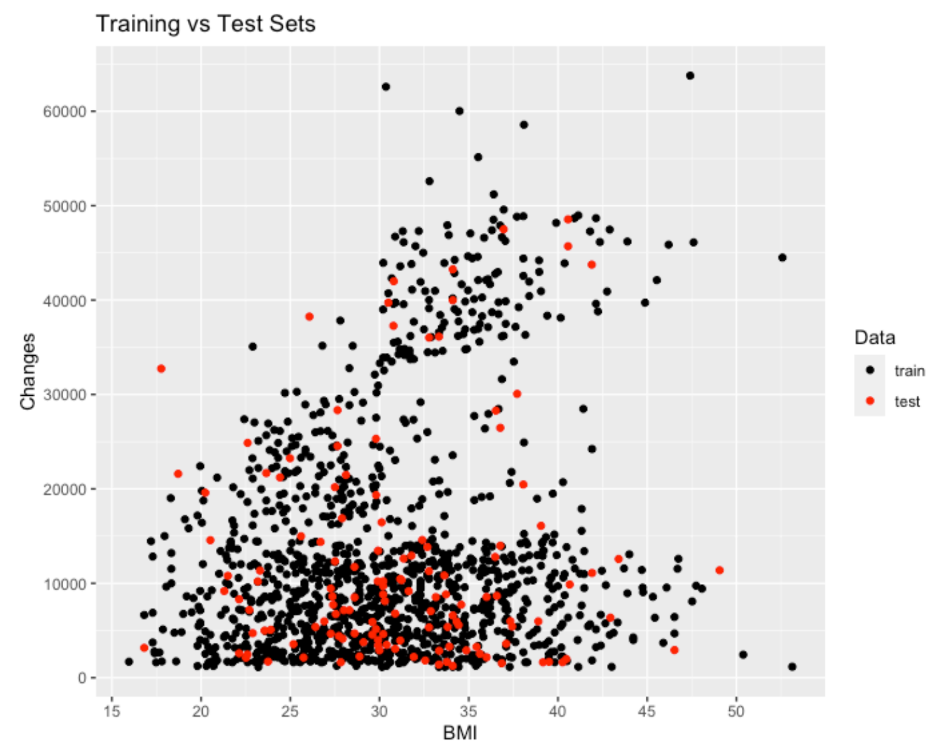
[12] Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1.

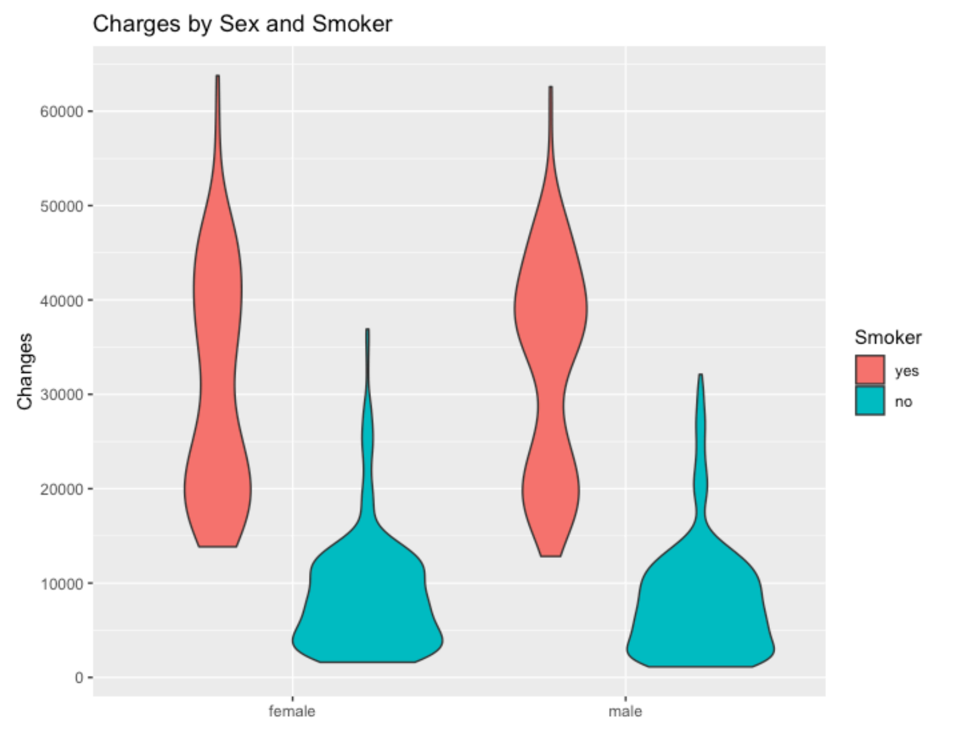
[13] Hahsler M, Piekenbrock M and Doran D (2019). "dbscan: Fast Density-Based Clustering with R.", Journal of Statistical Software, *91*(1), pp. 1-30. doi: 10.18637/jss.v091.i01

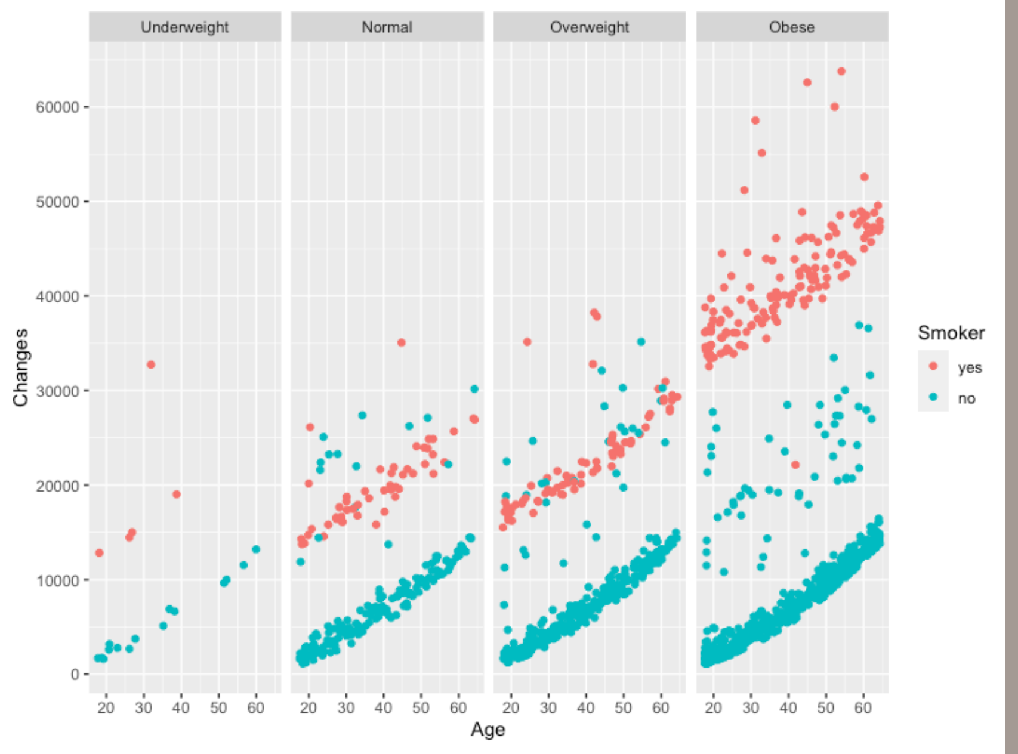
[14] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2018). caret: Classification and Regression Training. R package version 6.0-81.

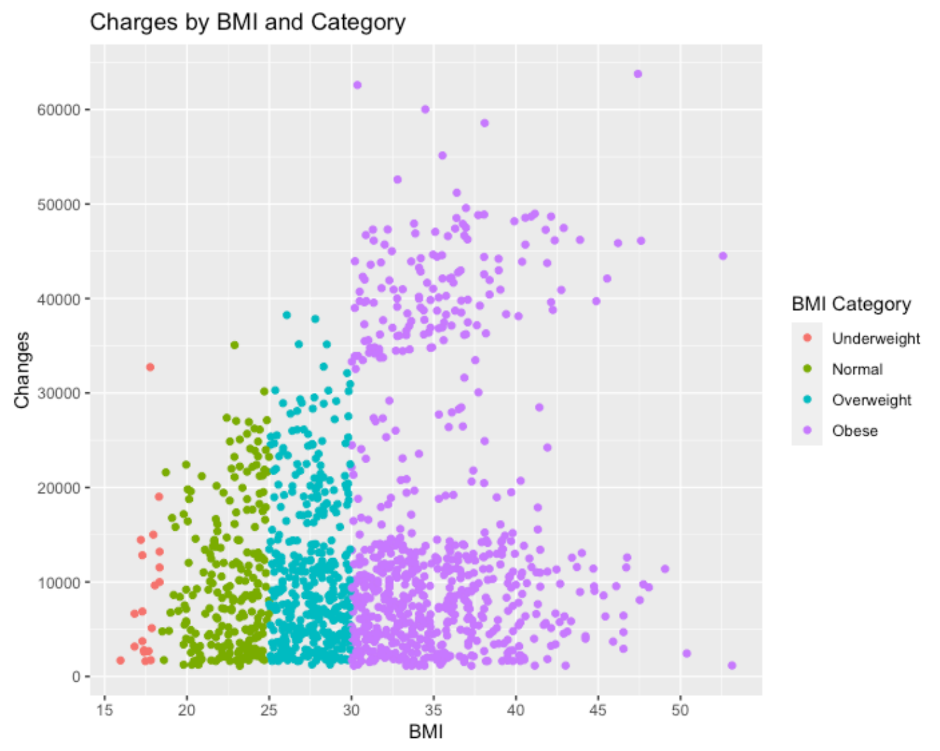
Results

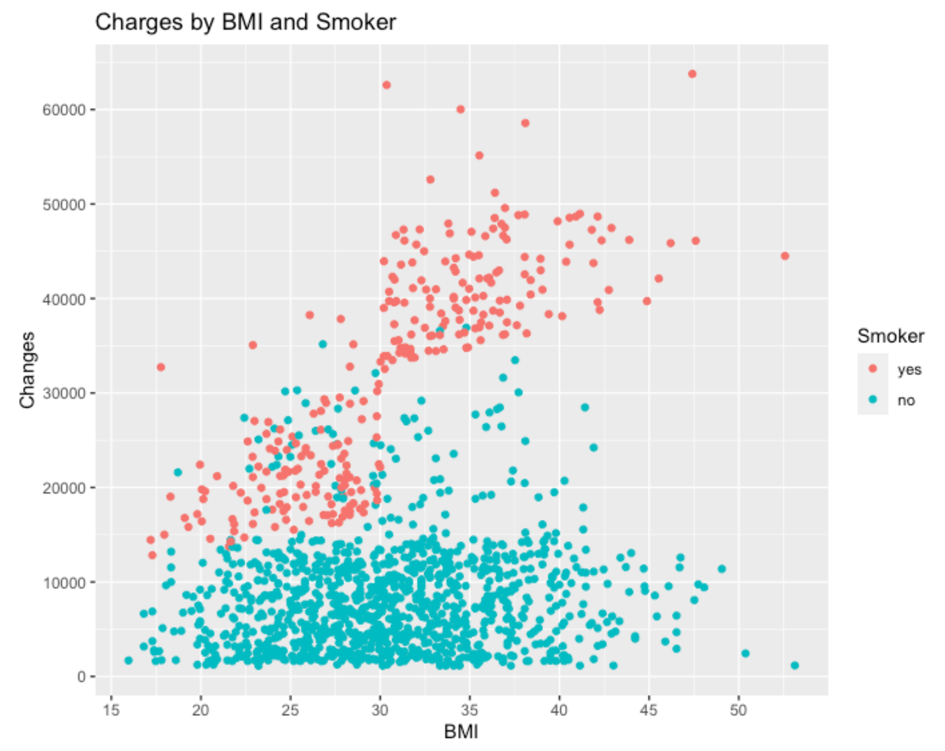


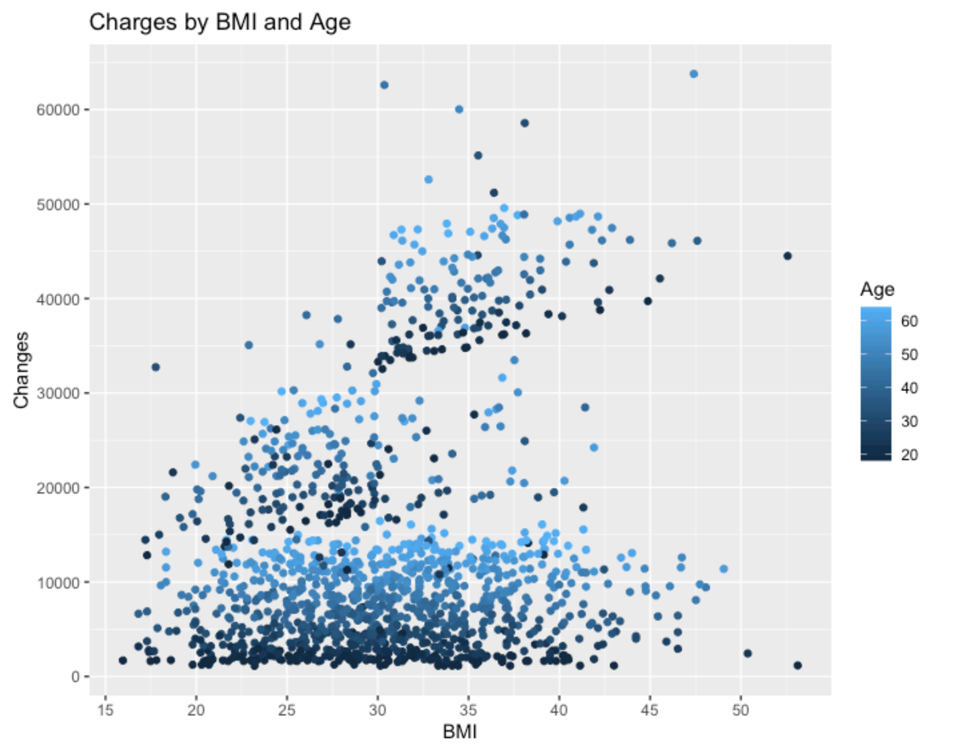


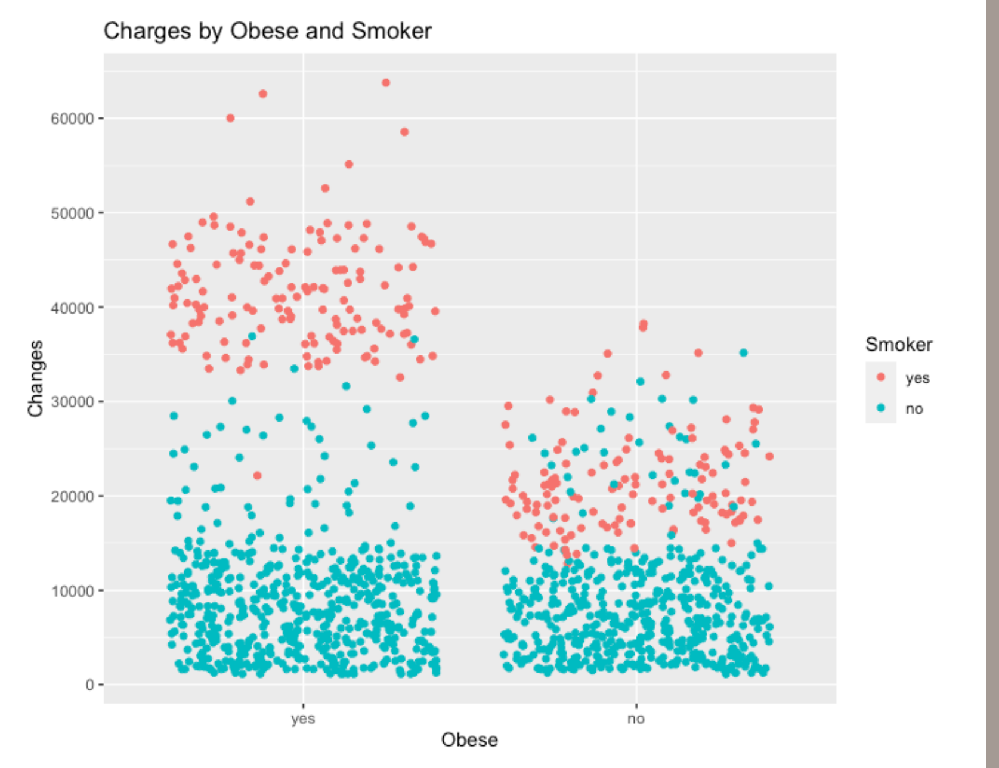


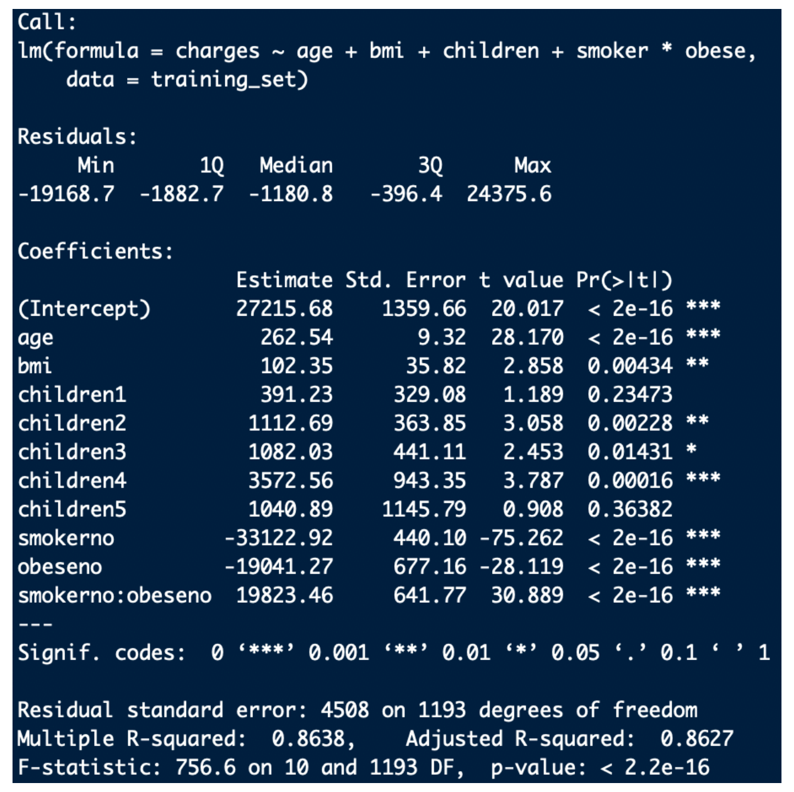




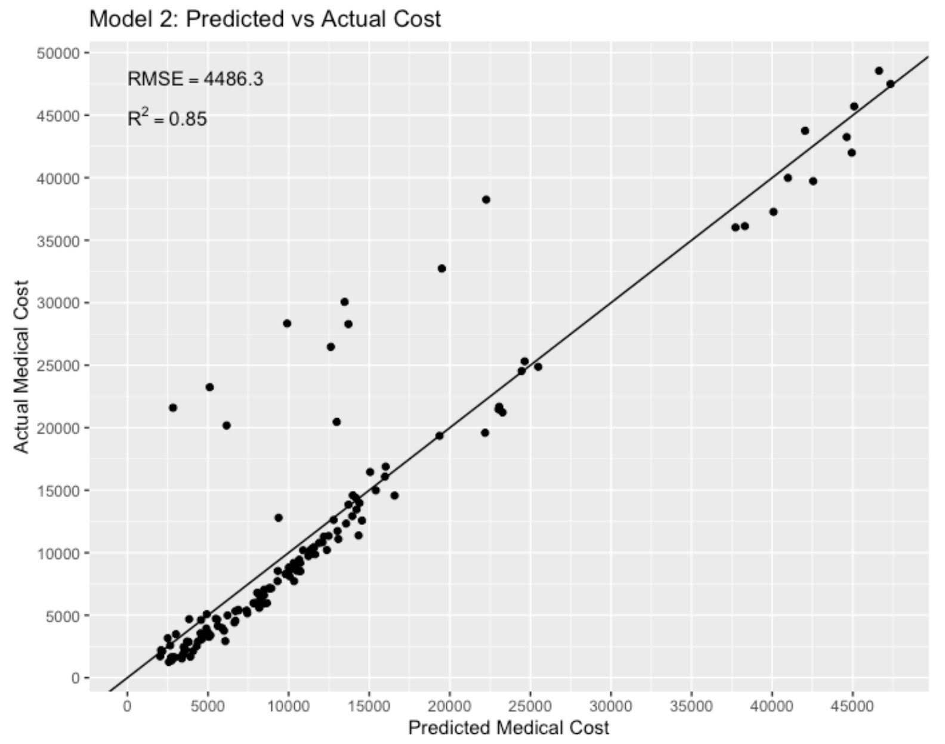


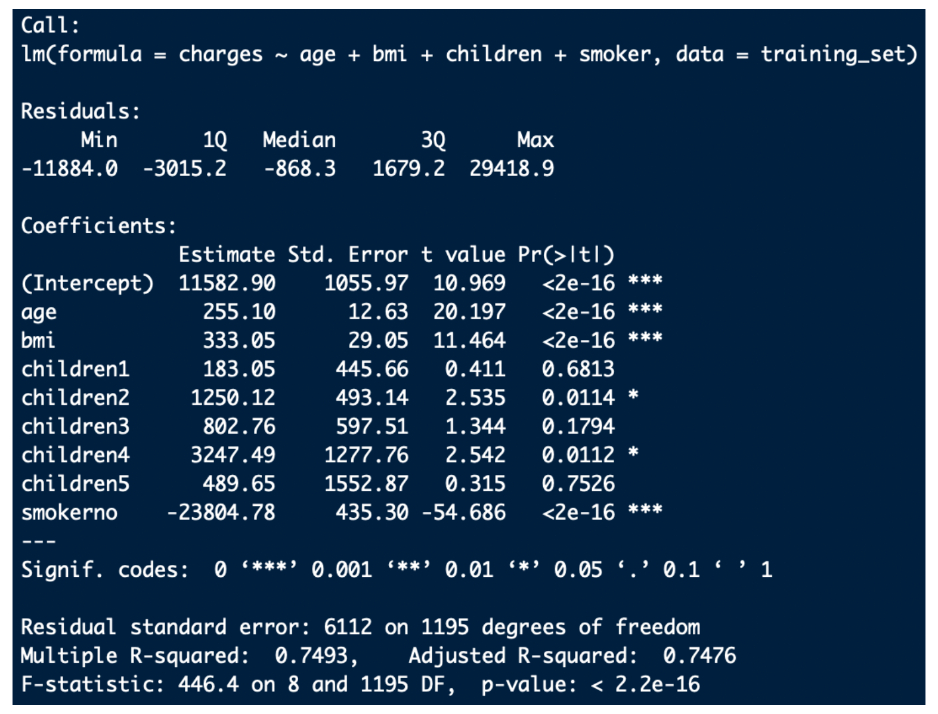




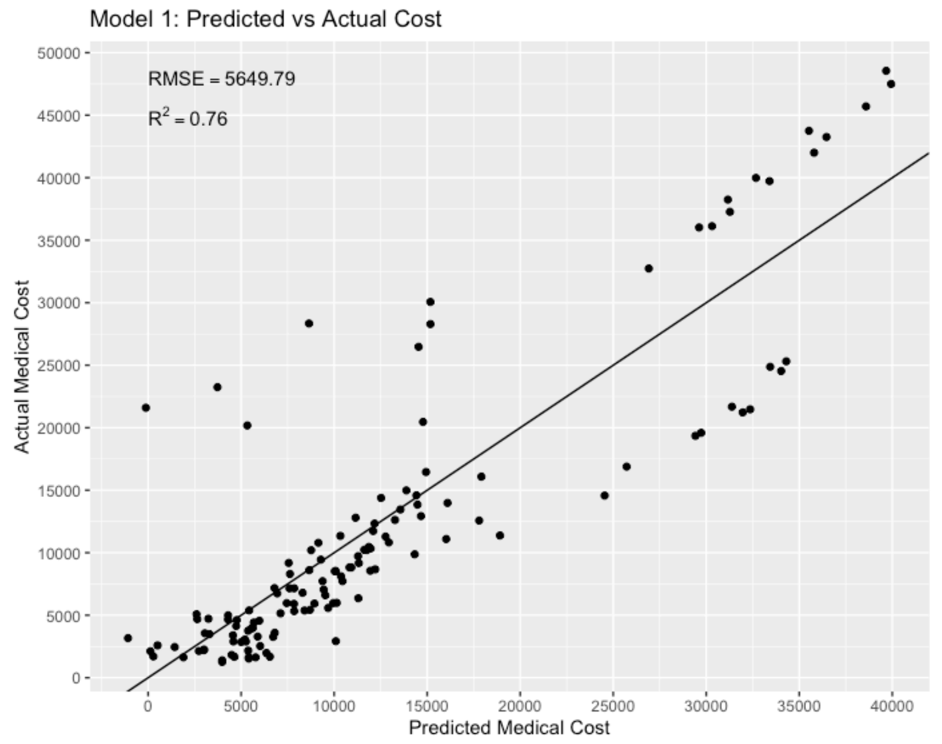


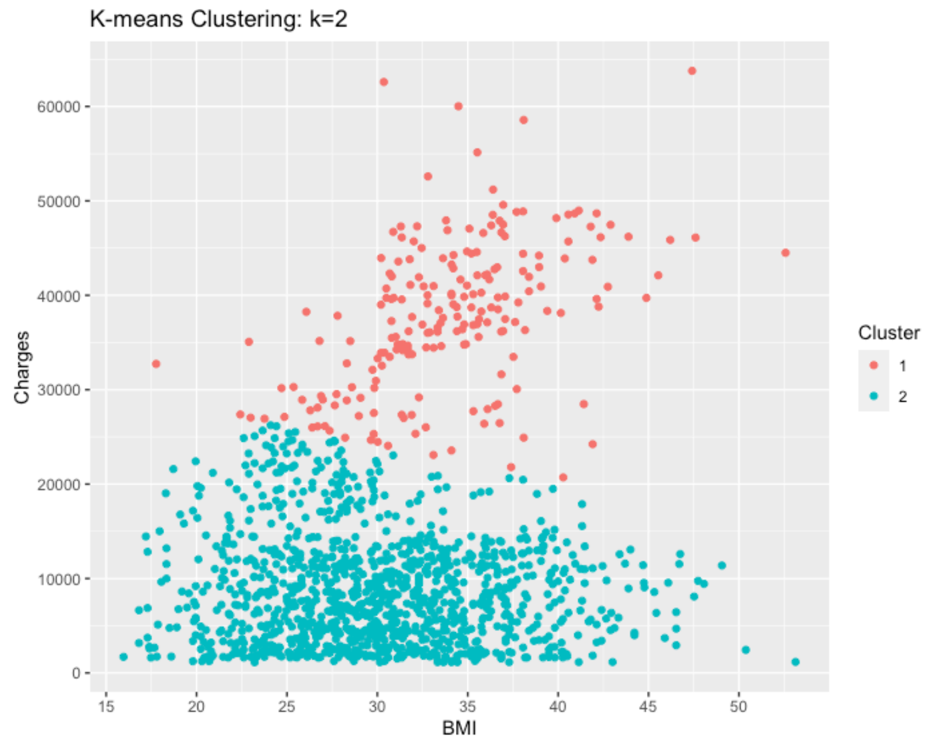
```
Call:
lm(formula = charges ~ age + bmi + children + smoker * obese,
    data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max
-19168.7  -1882.7  -1180.8   -396.4  24375.6

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       27215.68    1359.66  20.017  < 2e-16 ***
age                 262.54       9.32  28.170  < 2e-16 ***
bmi                 102.35      35.82   2.858  0.00434 **
children1           391.23     329.08   1.189  0.23473
children2          1112.69     363.85   3.058  0.00228 **
children3          1082.03     441.11   2.453  0.01431 *
children4          3572.56     943.35   3.787  0.00016 ***
children5          1040.89    1145.79   0.908  0.36382
smokerno         -33122.92     440.10 -75.262  < 2e-16 ***
obeseno          -19041.27     677.16 -28.119  < 2e-16 ***
smokerno:obeseno  19823.46     641.77  30.889  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4508 on 1193 degrees of freedom
Multiple R-squared:  0.8638,    Adjusted R-squared:  0.8627
F-statistic: 756.6 on 10 and 1193 DF,  p-value: < 2.2e-16
```

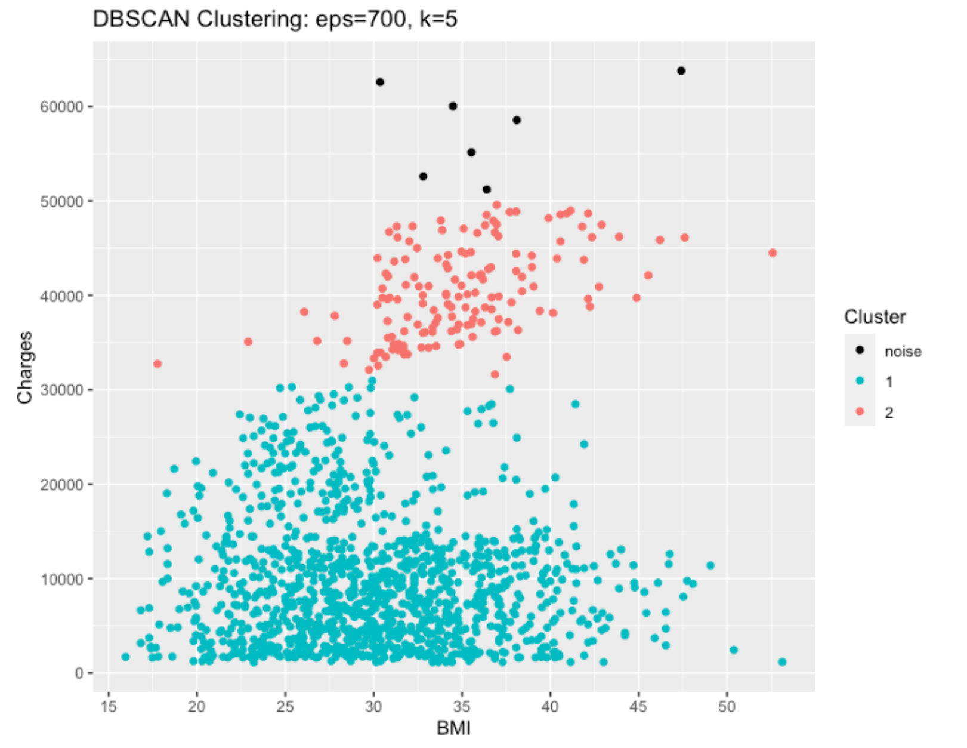


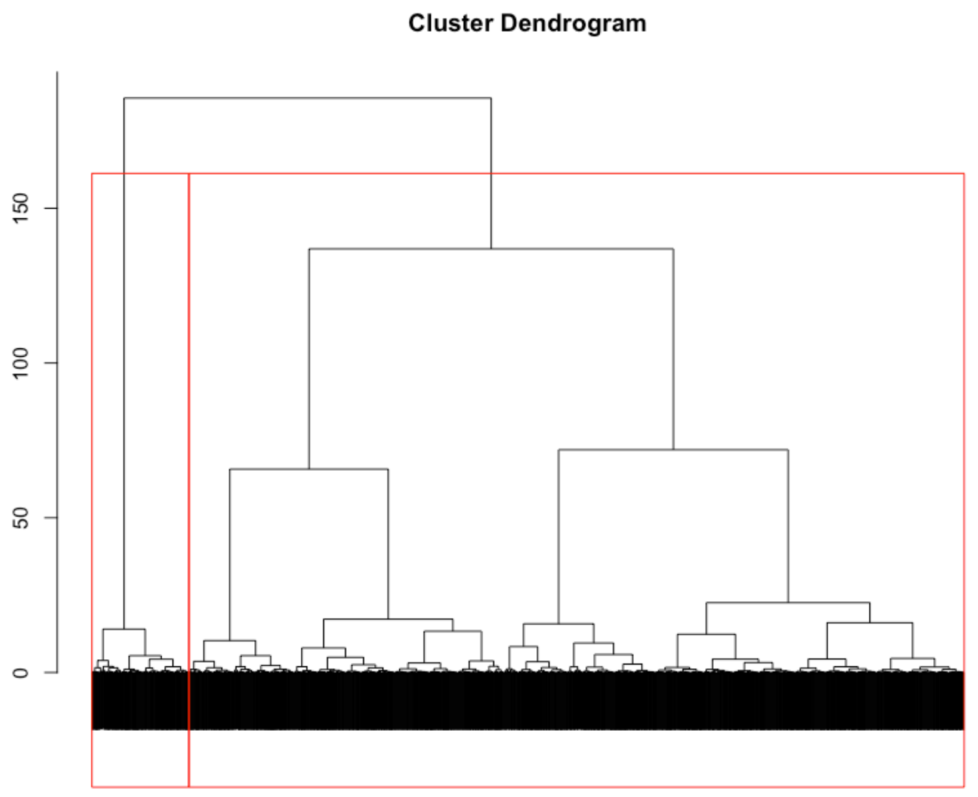


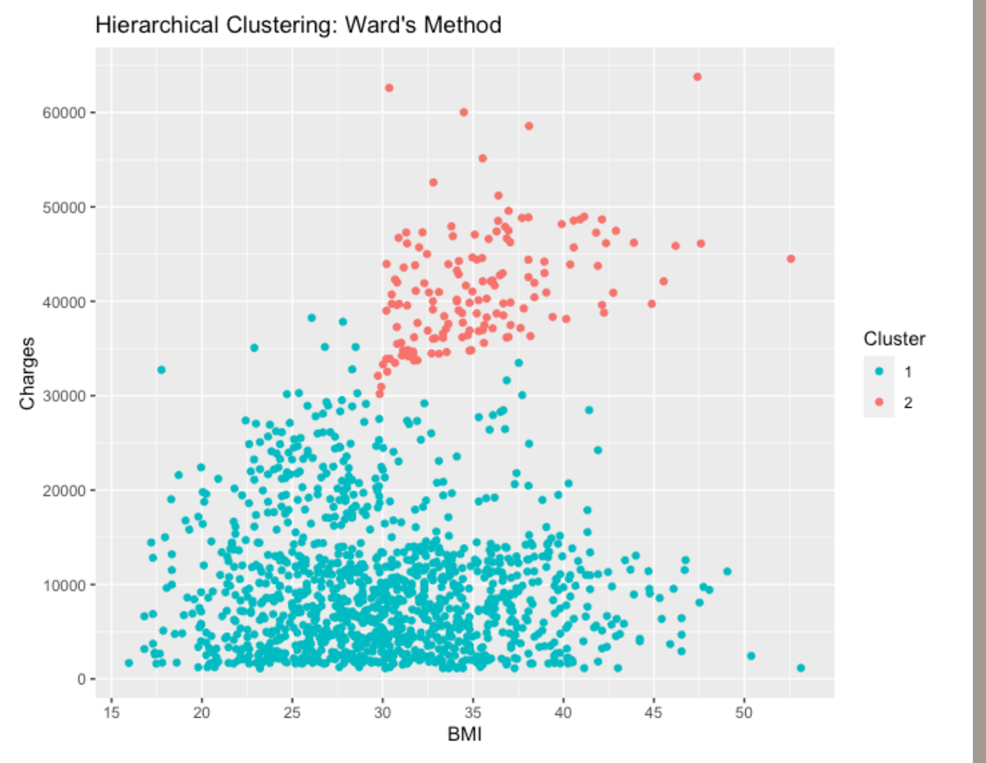
```
Call:
lm(formula = charges ~ age + bmi + children + smoker, data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max
-11884.0  -3015.2   -868.3   1679.2  29418.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11582.90    1055.97  10.969   <2e-16 ***
age           255.10      12.63  20.197   <2e-16 ***
bmi           333.05      29.05  11.464   <2e-16 ***
children1     183.05     445.66   0.411   0.6813
children2    1250.12     493.14   2.535   0.0114 *
children3     802.76     597.51   1.344   0.1794
children4    3247.49    1277.76   2.542   0.0112 *
children5     489.65    1552.87   0.315   0.7526
smokerno   -23804.78     435.30 -54.686   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6112 on 1195 degrees of freedom
Multiple R-squared:  0.7493,    Adjusted R-squared:  0.7476
F-statistic: 446.4 on 8 and 1195 DF,  p-value: < 2.2e-16
```


Model 1: Predicted vs Actual Cost
RMSE = 5649.79
R² = 0.76
Actual Medical Cost
Predicted Medical Cost


K-means Clustering: k=2
Charges
BMI
Cluster


DBSCAN Clustering: eps=700, k=5
Charges
BMI
Cluster
noise


Cluster Dendrogram
Height


Hierarchical Clustering: Ward's Method
Charges
BMI
Cluster