Can a Person’s Demographic Information Help Predict Insurance Costs?

Muhammad Banatwalla

December 7, 2020

1 Introduction

In recent years, U.S health care costs have increased exponentially. It ranks at the top of things Americans worry about and is a common topic of political discussion. The price of health care costs is often difficult to predict and unexpected medical bills can leave families in a bad financial state. In this study, we will be using data available on Kaggle from the U.S. Census Bureau on various demographic factors to predict individual medical costs billed by health insurance. The data set consists of 1338 observations and 7 variables describing an individual’s: age, sex, bmi, number of children, whether they smoke, the region where they live, and medical costs billed by their health insurance. To best determine this relationship, we will fit our data into various multiple regression models. The full model will allow us to test the inclusion/exclusion criteria of the individual variables.

1.1 Preliminary Treatment of the Data

First, we prepared the data for multiple linear regression by factorizing our categorical variables. Our collected data did not include any missing values. In order to find which variables should be used next, we checked for collinear variables. To get a general idea of which variables we should be looking at to test for collinearity, we created a correlation matrix.

	Age	Sex	BMI	Children	Smoker	Region	Charges
Age	1	0.021	0.109	0.0424	0.250	0.003	0.299
Sex	0.021	1	-0.0463	-0.017	0.076	-0.008	-0.057
BMI	0.109	-0.046	1	0.013	-0.004	0.157	0.198
Children	0.042	-0.017	0.013	1	-0.008	-0.001	0.068
Smoker	0.025	0.076	-0.004	-0.008	1	-0.013	-0.788
Region	0.003	-0.008	0.157	-0.002	-0.013	1	0.012
Charges	0.299	-0.057	0.198	0.068	-0.788	0.012	1

Table 1: Correlation Matrix of Variables

	GVIF
Age	1.0168
Sex	1.0089
BMI	1.1066
Children	1.0040
Smoker	1.0121
Region	1.0989

Table 2: VIF of Explanatory Variables

There are no obvious signs of multicollinearity, our strongest correlation is between charges and smoker at -.787. However, since the amount of charges is our response variable, we do not have any strong correlation among our predictor variables. Multicollinearity between predictors can give unreliable MLR results.We calculate the variation inflation factors (VIF) of our full model. GVIF is interpretable as the inflation in size of the confidence interval for the coefficients of the predictor variable in comparison with what would be obtained for uncorrelated data.Our results all indicate each variable’s VIF is close to 1. Again, we do not have any strong case for multicollinearity in our data.

2 Exploratory Data Analysis

We will employ data visualization techniques to find out more information about our data.

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

Table 3: Distribution of Charges

We see that the median of the data is below the mean and that the difference between mean and 1st quartile is greater than the difference between the mean and 3rd quartile. Both these facts indicate our data would be skewed to the right, We will verify with a histogram.

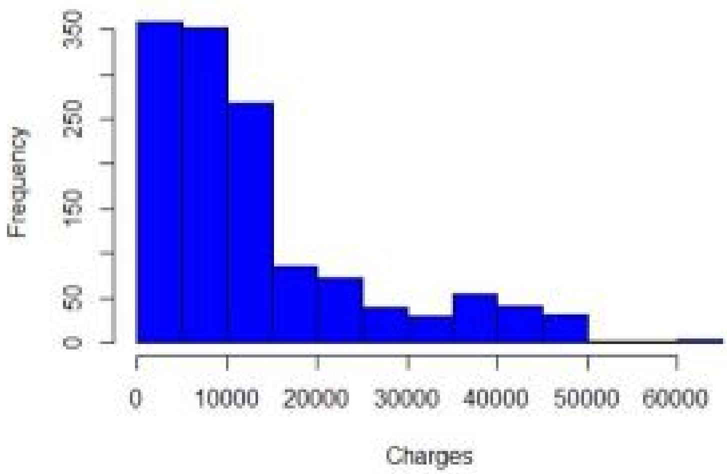

Figure 1: Distribution of Charges

Based on the histogram, the distribution of charges is skewed to the right, this offers us a challenge for our linear regression model.

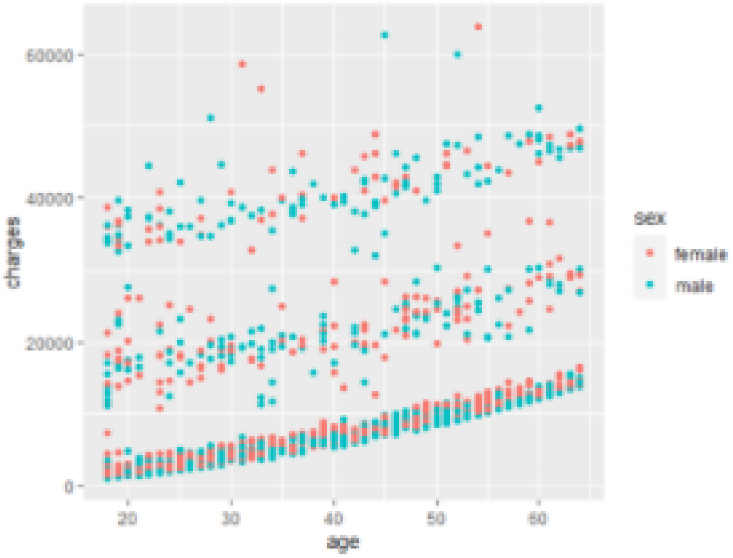

Figure 2: Age vs. Charges by Sex

We plotted age vs charges with color to denote their sex. According to the plot, we found no pattern between a person sex and insurance cost.

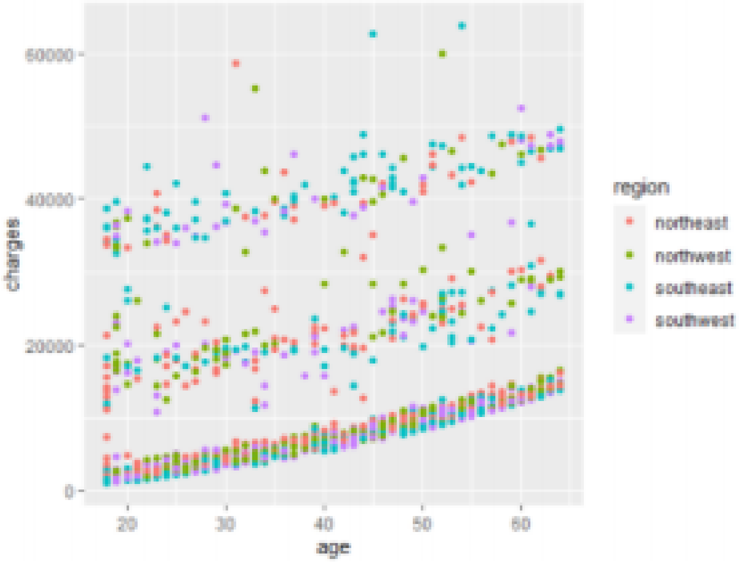

Figure 3: Age vs. Charges by Region

We graphed the same graph but this time colored the regions in different colors, We wanted to see if being from a particular region meant higher insurance costs. We found our data points to be well spread out when it comes to region and deduce that there must not be a strong correlation between one's region and insurance charges.

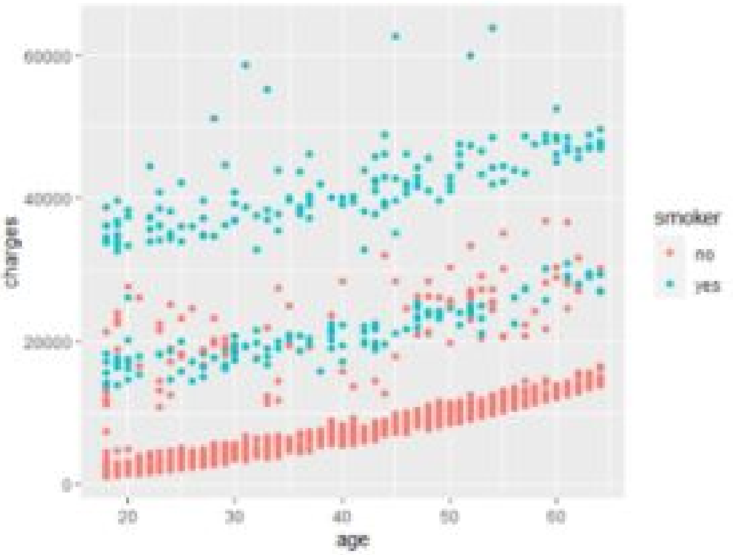

Figure 4: Age vs. Charges by Smoker

Since an individuals demographic information isn’t showing a strong correlation with charges, we hy pothesized that smoking would have a large affect on an individual’s medical charges. The data is split into three clusters, the lowest cluster is made up entirely of non-smokers and goes up slightly with increase in age. The second cluster is an almost equal mix of smokers and non-smokers and is more varied though it also tends to increase with age, the presence of this cluster shows that there might be another factor apart from smoking that might affect the insurance cost. The highest cluster is made up entirely of smokers, these are also people who get charged the highest from their insurance. This shows that smoking has a huge impact on insurance charges.

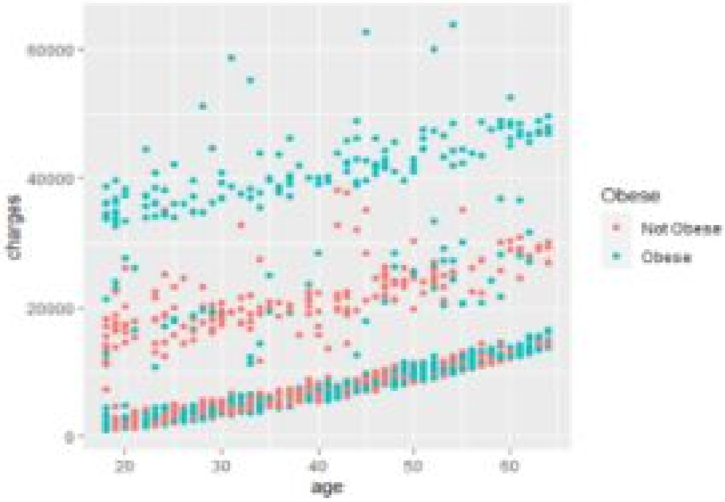

Figure 5: Age vs. Charges by Obese

We continued to explore our data this time we wanted to see the effect of obesity on insurance costs, we defined a person with a BMI greater then 30 to be obese and then ran our results. We found that the lowest clusters had some obese people in them. However, one interesting thing we found was that the people being charged the highest were almost exclusively defined as obese. Interestingly those people were strictly always smokers as well, this was interesting as it might seem there is a correlation between smoking and obesity.

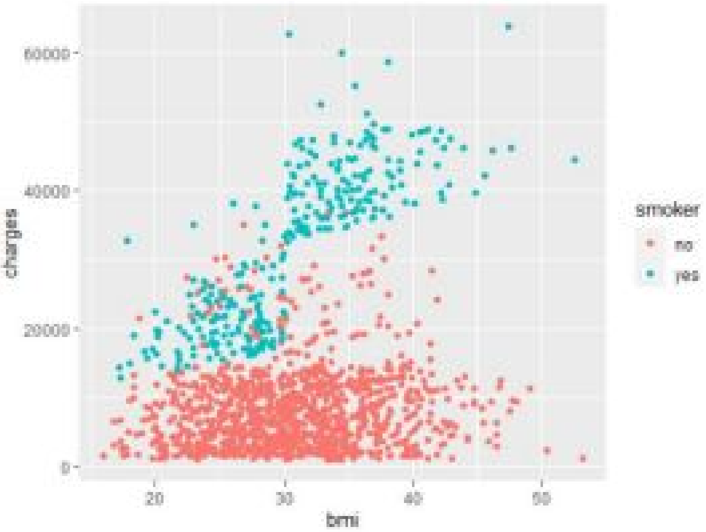

Figure 6: BMI vs. Charges by Smoker

To further investigate, we wished to see if having a higher BMI and also smoking had an impact on insurance cost. From our plot we can see that while insurance costs are higher for smokers in general, once the BMI increases from 30 the insurance cost gets significantly higher. We conclude that if you are both obese and smoking you will get extremely high insurance bills. This also confirm our results from figure 5.

3 Model 1: Charges vs. All Variables

Before we can even create a multiple linear regression model we need to test whether there is a regression relation through the general linear test, we used alpha = 0.01.
Hypothesis test: H_0: β_1, β_2, β_3,β_4,β_5,β_6 = 0 H_a: not all β_k equal zero
Decision Rule: F-statistic: 500.8 on 8 and 1329 DF, p-value: < 2.2e-16
Since the p-value is less than alpha 0.01, so we must reject the null hypothesis. We conclude that there is sufficient evidence at the 0.01 significance level that the insurance charges a patient receives is related to their age, bmi, sex, number of children they have, whether they smoke or not, and the region where they live.
The following is the regression equation for model 1:
Charges = - 11938.5 + 256.9*Age -131.3*SexMale + 339.2 *bmi + 475.5*children+23848.5*smokeryes - 353.0*regionnorthwest - 1035.0*regionsoutheast - 960.0*regionsouthwest

The first model contains all the predictors from the data set. We would like to select the best multiple linear regression model for our dataset that uses the best set of features to explain the most variation in the response. We did stepwise regression by backward elimination which deletes a variable at each stage to find the best model. The “best” model through stepwise regression keeps all the features. The criteria for the best model looks at the residual sum of squares and the Akaike information criterion (AIC) of each model, relative to each of the other models. The residual sum of squares (RSS) measures the amount of variance in a data set that is not explained by a regression model. AIC estimates the relative amount of information lost by a given model. This model was chosen as the best because it minimizes RSS and AIC. Now we need to make sure that our model is following the assumptions of multiple linear regression.

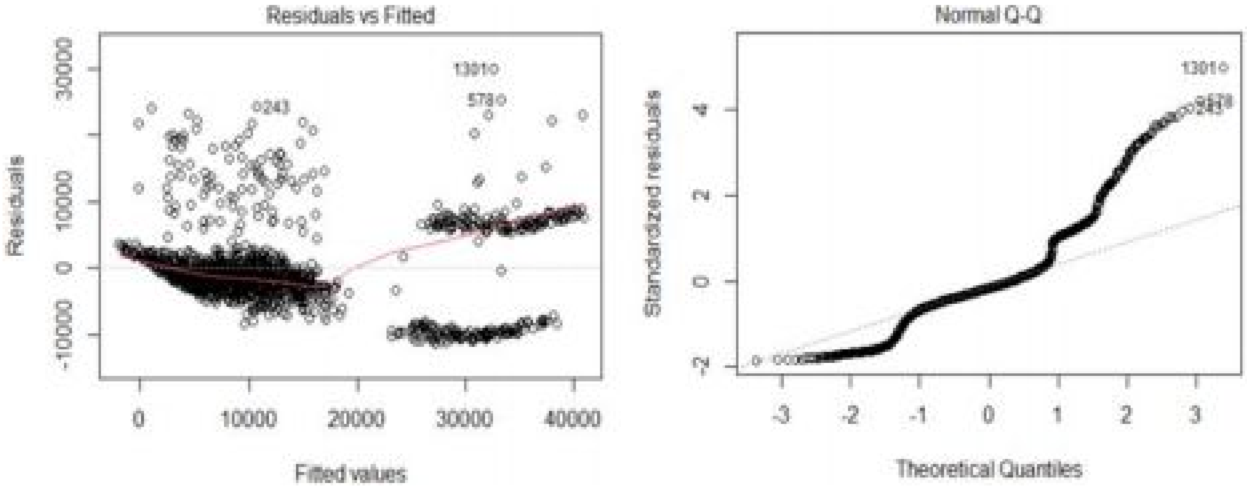

Figure 7: Residuals Plot and Q-Q Normal Plot of the Residuals of Model 1

The red line of the fitted values versus residuals plot is not flat, the linearity assumption is violated. The initial cluster of data points do hover around the 0 line, indicating equal variance, but after the gap the points stray further. The equal variance assumption is not met as a whole. The standardized residuals in our Q-Q plot do not follow the linear dotted line, therefore we conclude that the error terms are not normally distributed.

4 Model 2: Adding Polynomials, Binary Variables, and Interac tions

In our second model, we kept the original variables with the same dummy coding for sex, smoker, and region, but we decided to add age2, bmi30, and bmi30*smoker. After some research, our team discovered that age and medical charges do not have a linear relationship. Health insurance rates tend to increase disproportionately by age since companies predict that younger people would need less medical assistance and are generally less risky. For example, from age 21 to 25, the patient would pay a base rate of approximately $200, then the rate would go up about 10% every 5 years ($220 at age 30, $240 at age 35, etc.). However around age 50, the rates start to skyrocket at 1.80 times the base rate ($360) when it should be 1.50 times the base rate according to the linear pattern ($300). Therefore, we decided to make age2, the squared values of age, to resolve this non-linear relationship.

In terms of health insurance, the specific value of a person's body mass index is not important. It matters more whether the patient is in the range of underweight, normal weight, or obese. To simplify the calculations and because obesity is the greatest factor among the three on medical charges, we will focus on the obesity range. Obesity is considered to be a bmi over 30, so we created a binary indicator similar to our dummy variables. If the bmi is over 30 (obese), the new value is 1, and any bmi below 30 would convert to 0.

Our last addition is the interaction between patients that are both obese and a smoker. Our reasoning is that both of those variables are significant influences on charges alone, therefore, they should have even a larger impact on the dependent variable together. Our research revealed that smoking and obesity are closely linked. The risk of obesity and smoking are positively correlated and studies show that heavy smokers were more likely to be obese. The chances of obesity decreases the longer a former smoker abstains from smoking. The regression equation for model 2:

Charges = 69.2494 - 21.6786*age + 3.5978*age2 + 661.5105*children + 114.2920*bmi -475.6760*sexmale - 938.5116*bmi30 + 13421.6370*smokeryes - 275.6659*regionnorthwest - 826.1187*regionsoutheast - 1164.8152*regionsouthwest + 19912.6072*bmi30*smokeryes Now, we check if the assumptions of multiple linear regression are met of model 2.

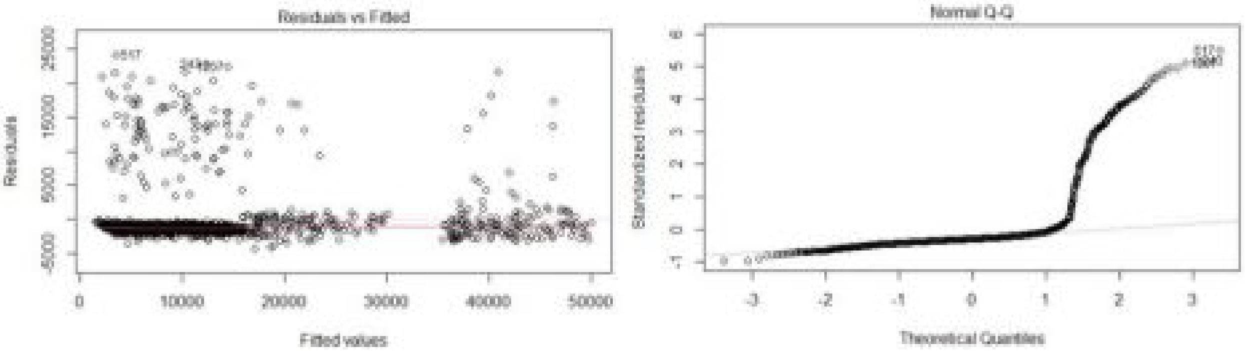

Figure 8: Residuals Plot and Q-Q Normal Plot of the Residuals of Model 2

Since the red line in the residual plot is fairly flat with no visible patterns, the linearity assumption is met. Overall, the majority of the data points demonstrate a ‘horizontal band’ around the 0 line, which indicated constant or equal variance. The outliers away from the band are likely not influential since they all follow the x value trend and will not displace the fitted function substantially. The Q-Q plot of the residuals demonstrates that our residuals roughly follow a normal distribution. In the Q-Q plot, the initial residual points follow the dotted line, then it has a steep increase after the first quartile. The vertical jump suggests that there is a density gap in the sample, which is also shown in our initial residual plot.

5 Discussion

In order to determine the “best” model we need to analyze the R2 value and adjusted R2 value of both models.

Model	R	Adj R	RMSE
Model 1	0.7509	0.7494	6041.68
Model 2	0.8679	0.8671	4399.45

Table 4: R-squared, Adjusted R-squared, and RMSE for both models

For model 1, according to the coefficient of determination (R^2) 75% of the variation in insurance charges is explained by the variation in the entire set of x variables. The R^2 value for model 2 is much higher, indicating that model 2 better explains the variation in our data. The adjusted R^2 only takes into account the independent variables that actually affect the dependent variable, and subsequently, penalizes if unrelated variables are in the model. We can see the model is not over-fitting the data for either of the models because the difference between adjusted R^2 value and R^2 value is quite small.The adjusted $R^2$for model 2 is also higher at .8671. An increased adjusted coefficient of determination supports that Model 2 is a better fit than Model 1. The root-mean-squared-error measures the predicted values of the model and the actual values ideally we would want it to be close to zero. Model 2 has a lower RMSE verifying that it is the better model.

To evaluate the performance our best model, we randomly split our data choosing 80% of our data to train our model and leaving 20% to test and evaluate its performance. Since in any machine learning model we want to capture the true relationship between the predictors and the response, we want little bias. We also want low variance, if we use new data that the model has never seen before we want to have similar performance. This is the benefit of train-test-split we can evaluate the model’s performance on data it has never seen before. We made predictions for the charges on the test set using model 2 then plotted it against the true charges for these patients. This allows us to visualize how well our model predicts insurance costs.

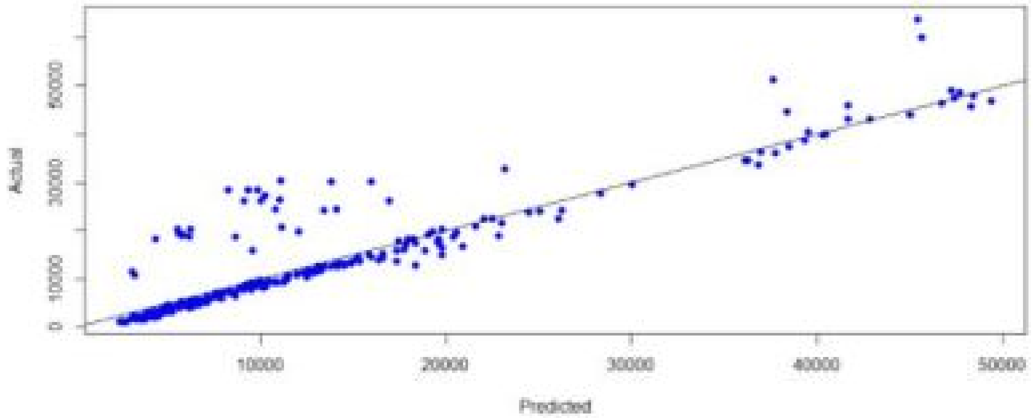

Figure 9: Predicted Values vs. Actual Values

According to figure 6, the points are nearly linear; this indicates good predictions. However, there are some points away from the trend. To further investigate we randomly chose five patients and their predicted insurance costs from model 2 and their true costs.

Patient	Age	Sex	BMI	Children	Smoke	Region	Pred	True
case #5	32	M	28.88	0	No	NW	$5,131.88	$3866.86
case #12	62	F	26.29	0	Yes	SE	$28,319.25	$27808.73
case #225	42	M	24.64	0	Yes	SE	$20,580.19	$19,515.54
case #506	37	M	30.88	3	No	NW	$7,685.34	$6,796.86
case #550	43	F	46.2	0	Yes	SE	$41,673.40	$45,863.21

Table 5: 5 Randomly Chosen Patients from the Data, their Characteristics, True Prices, and Predicted Prices using Model 2.

We can see that Model 2 has a high accuracy of prediction for some cases and low accuracy for others. Four of the randomly chosen cases have predicted costs less than $1200 from true costs. The last case has a predicted price $4000 away. The model represents 87% of the variation in the data. The model may have had a harder time representing this patient.

6 Conclusion

The main goal of this project was to determine how the charges billed by insurance provider’s vary based on several characteristics of each patient. Initially, we did a preliminary analysis of the data and created a multiple linear regression model with all the features in the data set. Through further analysis, we determined that a patient’s body mass index and whether they smoked or not played a big role on the medical charges they received. With this information we created a new model that explained more of the variation in charges billed to patients and made predictions of these charges. The model predicted well for some patients and not others. One of the limitations of our data set was that charges, the response variable, had more patients with lower charges. This could have caused the model to have difficulty predicting patients that had higher charges since it didn’t have as many cases to learn from. Overall, we learned that behavioral factors play a bigger role on medical charges than their demographic information. This information can be useful to many people whom unexpected medical bills will put them in a bad place financially.

7 Appendix

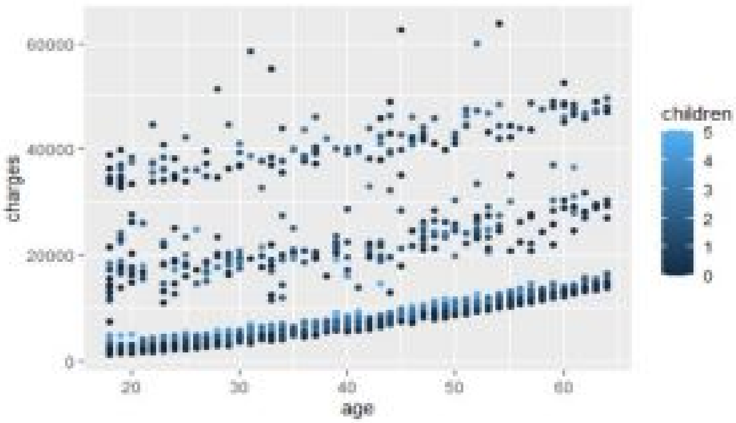

Figure 10: Age vs. Charges by Number of Children

We were wondering if there is a relationship between the number of children and insurance charges, we thought out middle cluster would contain point with most children but the graph proved us wrong, we can not say that the number of children has an impact on the insurance cost.

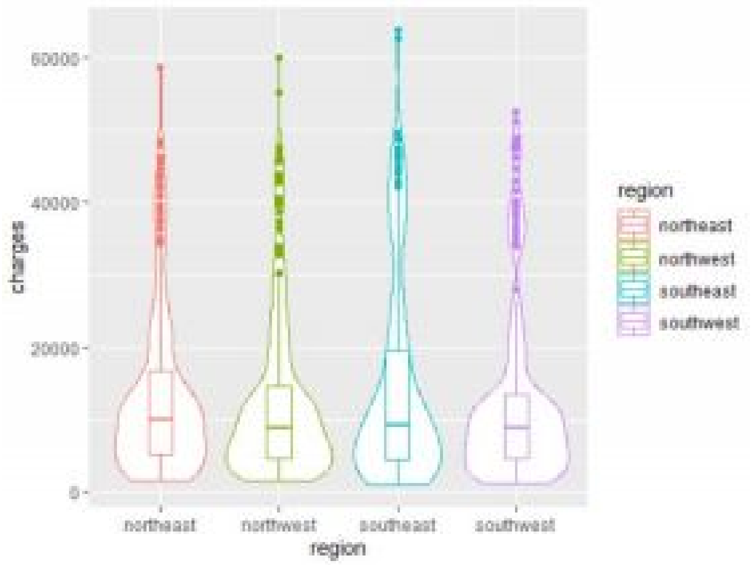

Figure 11: Region vs. Charges

We plotted boxplots inside violin graphs to see the distribution of our data separated by region, we noticed the distribution to be identical with identical means, the distribution for southeast is the most varied and the distribution for the southwest in the least varied. There are several outliers for all the regions we assume these outliers to be smokers and obese people.

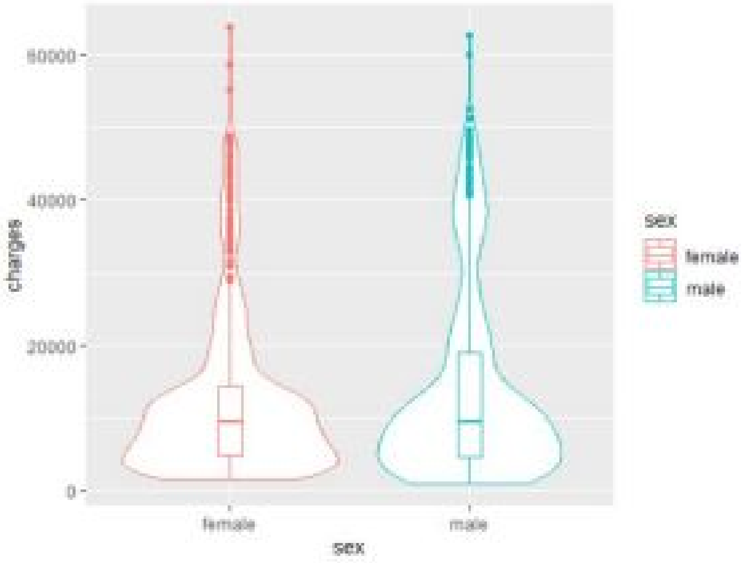

Figure 12: Sex vs. Charges

We drew boxplots inside violin graphs to see the distribution of our data separated by sex, we noticed the distribution to be identical with identical means. Both male and female have low means for charges and majority of the distribution lies beneath $20,000 annually. There are several outliers for both sexes and again we believe they are due to smoking and obesity.

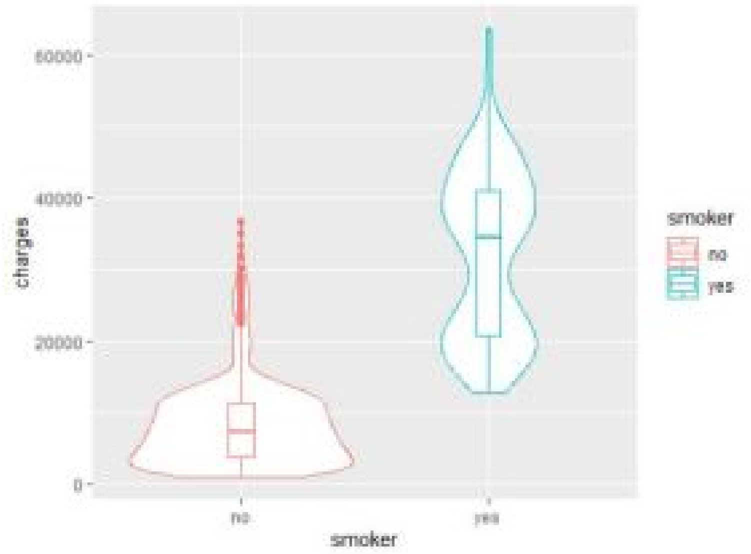

Figure 13: Smoker vs. Charges

We drew boxplots inside violin graphs to see the distribution of our data separated by whether a person smokes or not. As expected there is a clear difference between the distributions of those who smoke and those that don’t smoke. The mean of those who don’t smoke is way less than those who do. One thing of note is that the yes violin plot shows two humps which means that there are two cluster fro those who smoke, we believe the higher cluster is probably peopl who are also obese. We have some outliers for people who dont smoke, we think they might be the people who are obese.

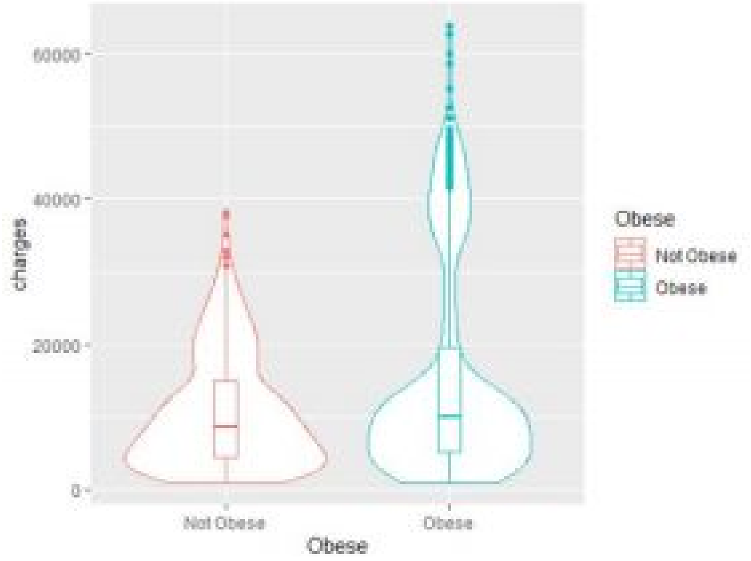

Figure 14: Obese vs. Charges

We drew boxplots inside violin graphs to see the distribution of our data separated by whether a person is obese or not. The mean of both obese and not obese is about the same and for both a vast majority of the data lies below $30,000 anually. The difference however is that the distribution of obese cases is more varied and there are a lot of cases higher then the 3rd quartile, again these are probably the smokers.

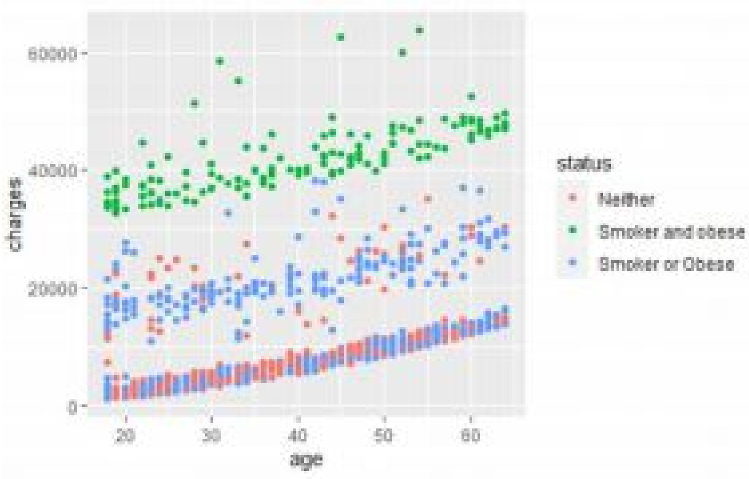

Figure 15: Age vs. Charges by Smoking and Obese Status

We continued on our quest to find out what makes up our second cluster. We computed a new variable that checks whether a person is both “A smoker and Obese”, either “A Smoker” or “obese” , or “neither”. Our results show that the top most cluster is made of people who are both smokers and obese, the middle cluster is made up of mostly the people who are either one or the other and the bottom cluster is made up of people who are either obese and non-smokers or are neither obese or smokers.