Term Project Report

Team DWLS

Devaney Loeza | Wayne Chim | Lamia Aesha | Steven Walker

Introduction

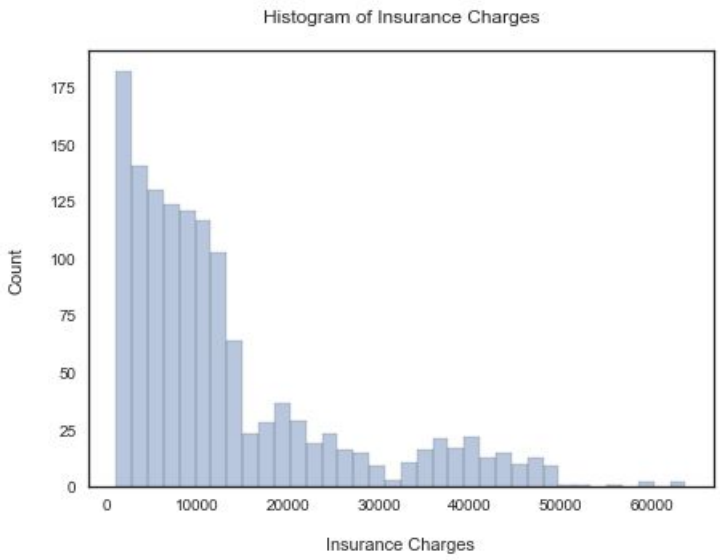
This dataset is intended to be used to develop a model to predict insurance charges based on several parameters. There are 1338 records in this data set with 7 variables. The records consist of 49.5% females and 50.5% males. 79.5% are non-smokers and 20.5% are smokers.

Predictors	*Description*
Age	*Age in years*
Sex	*1 = Male; 0 = Female*
BMI	*Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9*
Children	*Number of children*
Smoker	*1 = Smoker; 0 = Non-smoker*
Region	*Residential area in the U.S.* *Northeast, Northwest, Southeast, Southwest*
Charges	*Individual medical costs billed by health insurance*

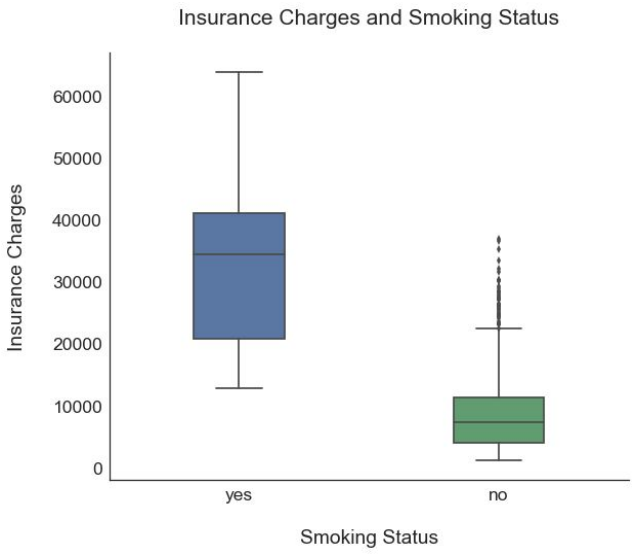
This dataset is intended for a general audience; it can be beneficial for anyone to know what factors go into rising medical costs. This data could also be useful to insurance companies, so they know how much to charge someone based on certain predictors.

EDA Models

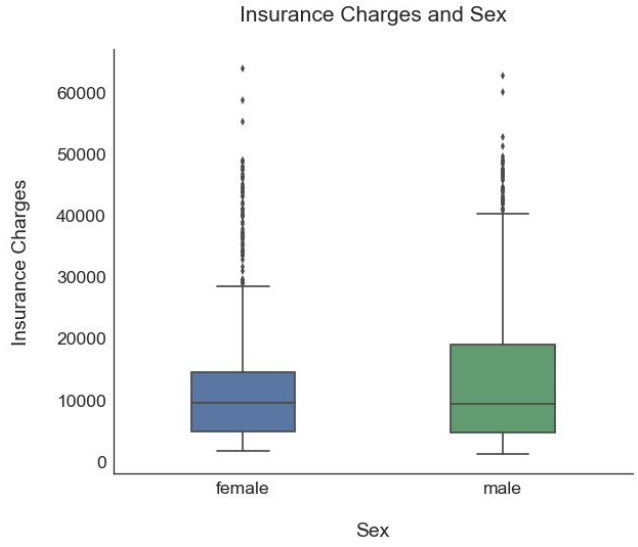
A majority of insurance charges are under $30,000, only a small portion go beyond that price range. We can observe that highest count rate for charges is lower than $5,000. And majority of the data falls underneath $20,000. We have a couple charges that seem to be outliers of $60,000 plus.



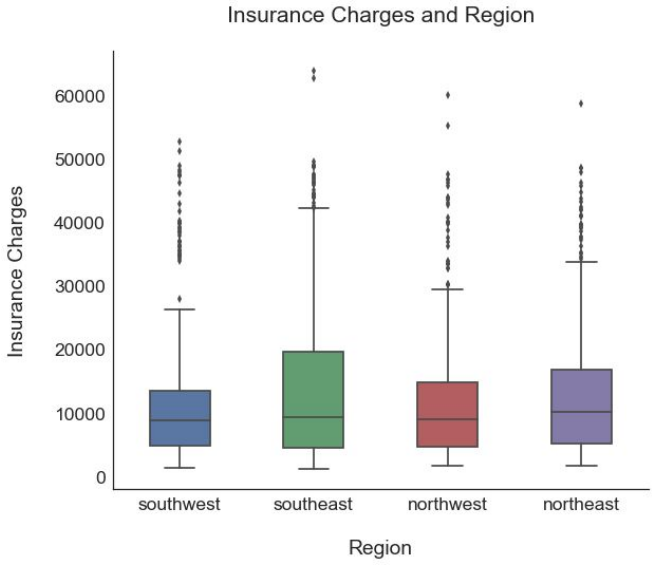
Taking account of the predictor *Smoker*, it indicates a greater average in insurance charges for smokers in comparison to non-smokers. We can also see a couple of outliers in both categories. We have a median charge of a little over $30,000 for smokers and a median charge of a little under $10,000 for non-smoker. That is a $20,000 difference on average between smokers and non-smokers. This gives us a clear view on what we are looking for when we create our predictive model using the data.



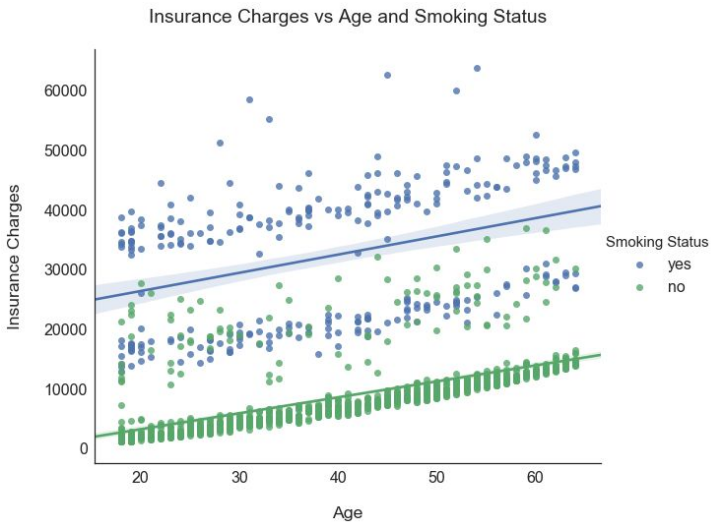
This boxplot shows insurance charges separated by sex. The distribution is similar for both sexes, with males accumulating slightly more insurance charges than females. This could possibly give us a clue the strength of sex on our predictive model. We can also view some outliers between male and females. Both sexes have outliers which gives us a hint that the outliers in the dataset is also pretty evenly distributed between sexes.



Insurance charges were similar across each region, with the Southeastern U.S. reporting slightly higher charges. Questions we want to ask ourselves would be; does location really play a significant part in insurance charges in the U.S.



This scatterplot shows how insurance charges relate to age. Two regression lines are shown: one for smokers (blue) and one for non-smokers (green). This visual shows that insurance charges increase the older someone is; it also shows that smokers accumulate more medical costs than non-smokers.



Linear Regression Model Development

Charges, the response variable, is continuous, so the regression model would be linear. The first model takes all predictors into account. The summary shows that *Age*, *BMI*, *Children*, *Smoker*, and *Region* are statistically significant, which eliminates consideration of *Sex* in the following models. On another note, *Region* has the highest p-value at 0.064, making it the least significant of the five predictors. The second model is built with the most statistically significant predictors, *Age*, *BMI*, *Children*, and *Smoker*. The third model is more or less the same as the former, but with the addition of *Region*. It would be interesting to evaluate *Region*'s impact on model quality and performance.

Model 1	$Charges = \beta 1*Age+\beta 2*Sex+\beta 3*BMI+\beta 4*Children+\beta 5*Smoker+\beta 6*Region+\beta 0+\epsilon$
Model 2	$Charges = \beta 1*Age+\beta 2*BMI+\beta 3*Children+\beta 4*Smoker+\beta 0+\epsilon$
Model 3	$Charges = \beta 1*Age+\beta 2*BMI+\beta 3*Children+\beta 4*Smoker+\beta 5*Region+\beta 0+\epsilon$

AIC, BIC, and Adjusted R-squared for Model Comparison

Model 3 is evaluated as the better model in terms of AIC and Adjusted R-squared, however the magnitude of the difference in the two metrics for the three models is miniscule compared to the difference across BIC. Compared to AIC, BIC is more sensitive to parameters, which pertains to actual models, and therefore is a more consistent metric. More weight is placed onto the results in BIC, and the better model is Model 2.

	AIC	BIC	Adjusted R-squared
Model 1	27114.44	27156.03	0.7492
Model 2	27114.04	27145.23	0.7489
Model 3	27112.59	27148.98	0.7494

Cross Validation for Model Comparison

A ten-fold cross validation resulted that Models 2 and 3 performs significantly better than Model 1. Between the two, Model 3 has a slight edge over Model 2, however more tests are required before selecting the ideal predictive model.

	10-Fold CV
Model 1	36890854.48
Model 2	36874298.05
Model 3	36872496.50

Stepwise Selection

When creating a predictive model we want to select the most significant variables. Using the stepwise selection method we can see which variables have a significant impact on the dependent variable; Charges. There are different ways to check the weight of the variables. The different types are listed below for reference.

Backwards Selection - starts off with all predictors in the full model, then interactively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.
Forward Selection - starts with no predictors in the model, then adds the most contributive predictors.
Hybrid Selection - which is a combination of forward and backward selections. You start with no predictors, then s add the most contributive predictors. After adding each new variable, remove any variables that no longer provide an improvement in the model fit

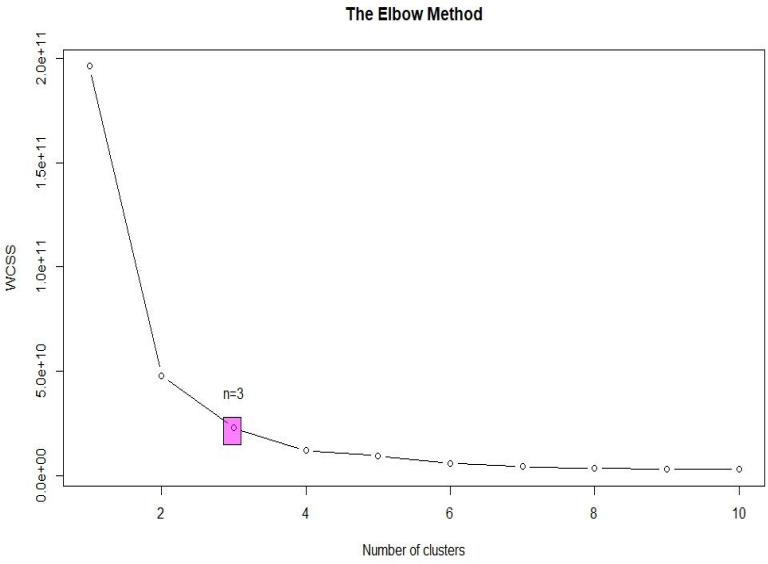
When running this methods on R we can conclude that certain combinations of variables do make a statistical significant on *Charges* upon analysis. We can see the variable combinations of Model 1, Model 3, and Model 2. We can see that all six variables have a weight on charges, but in different combinations. The first variable that we can see has a high significant is *Smoker*, then the third combination is *Age*, *BMI*, and *Smoker*. The fourth combination created was *Age*, *BMI*, *Children*, and *Smoker*. The fifth combination is *Age*, *BMI*, *Children, Region* and *Smoker*. The last combination was a result of all the variables.

Best Subset Selection

Now that we have found the different combinations of variables that have significant statistically components to charges. Using Best Subset Selection we can see which variables have a high significance. This technique is used to reinforce the stepwise selection technique. When carrying this out we can see that *Age* and *Smoker* have a high contribution to insurance charges. But just using these two variables alone to construct the final model will not be enough to have accurate model for predictive purposes. Just using two variables (*Age* and *Smoker*) were used to construct Model 4. Model 4 contained a higher AIC and BIC then the other models. We can see from this demonstrative model that only having the variables *Age* and *Smoker* to predict

Charges generates a higher AIC of 27,253.32 and a BIC of 27,274.12. Then the other models. Concluding that using this technique was helpful in fully understanding the measure of each variable and how they play a role in measuring *Charges.*

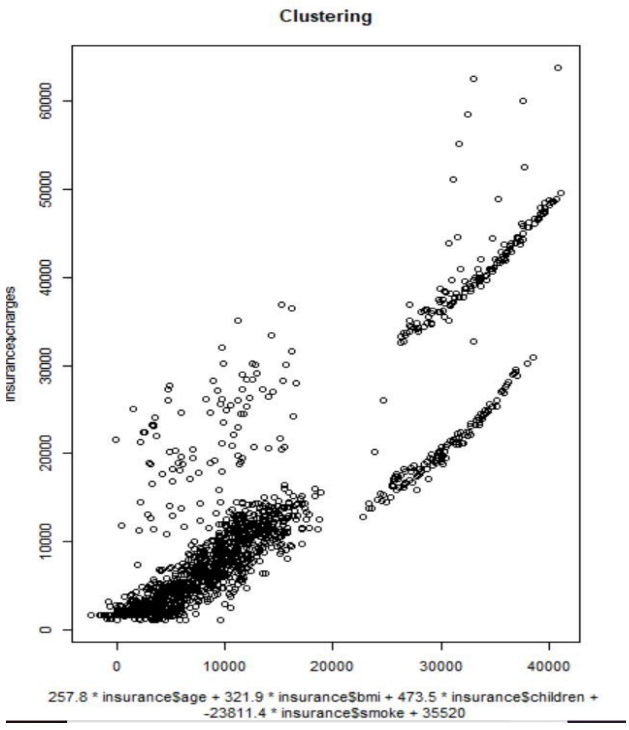
Hierarchical Clustering



The elbow method revealed that the appropriate number of clusters for the model is 3, which also indicates that there are 3 homogeneous variables for predicting insurance charges.

By partitioning the regression into 3 clusters: the first cluster contain people with insurance charges with $20,000 or less, the second cluster contain insurance charges in the range of $10,000 to $29,999, and the third cluster has the most expensive charges at greater than $30,000.

The values for *BMI* at all clusters were evenly distributed, so it alludes that *Age*, *Children*, and *Smoker* has a more significant impact on *Charges* and would reveal certain patterns. Similar to BMI, age also had a relatively even distribution among the clusters. However, once age was categorized into ranges, it showed that as charges went higher, the age range in majority grew older with it -- going from 18-33 in the first cluster to 51-64 in the third cluster. Surprisingly, the amount of people with 0 children remained at large, hovering around 40~47% in all three clusters. As insurance charges went higher, the percentage for other number of children in the family grew from approximately 14% having a child in the first cluster to 45% having 1-2 children in the third cluster. Finally, as expected, a smoking habit has the most dramatic statistics within the clusters. The smallest charges in the first cluster amounted to 94.17% non-smokers, dropping slightly to 73.7% in the slightly more expensive



charges in the second cluster, and arriving at a staggering 93.8% smokers in the most expensive charges.

Research Comparison

After concluding that Model 2 was the better model we decided to put our model to the test, in a case study. We wanted to see the differences in Charges against the other variables. As we stated Age and Smoker have a high impact on Charges. We decided to see the difference of a male with the same BMI , 0 children and the same ages (26) to see the contrast between smoker and non-smoker made on his charges. The results were very high. We could see that a non-smokers charge was $5,611.82 while a smokers charge was $29,423.22. We also looked into the how age weighted a difference on charges. We saw the same results of a 56 year old smoker with 0 children at $37,158.72 and a 56 year old non-smoker at $13,347.32. Adding children into these case studies only slightly change the value of charges. Still children did have an impact on charges.

Conclusion

Charges = 257.85*Age + 321.85*BMI + 473.50*Children - 23811.40*Smoker + 35520.03

Model 2 is the better regression model. *Smoker* is the predictor with the most profound impact on insurance charges. As a categorical variable, smokers are assigned a value of 1 while non-smokers are assigned a value of 2. In other words, a smoking habit can potentially increase insurance charges by $23,811.40.

In addition to developing an accurate predictive model for insurance charges, it is interesting to mathematically and statistically prove the devastating financial impact a smoking habit has beyond health concerns.