**predicting medical expenses using MULTIPLE regression**

3251-014

CHAEWON KIM

AUGUST 2018

# TABLE OF CONTENTS

[TABLE OF CONTENTS 2](#_Toc521415918)

[Objective 3](#_Toc521415919)

[What are you setting out to prove or predict? 3](#_Toc521415920)

[What is your rationale for there being a correlation in the data that you’re looking to confirm and/or exploit? 3](#_Toc521415921)

[Data Preparation: 3](#_Toc521415922)

[What was your data source? How good was the data quality? 3](#_Toc521415923)

[What did you need to do to procure it? What tools or code did you need to use to prepare it for analysis? 3](#_Toc521415924)

[What challenges did you face? 3](#_Toc521415925)

[Analysis 4](#_Toc521415926)

[Conclusion 10](#_Toc521415927)

[Appendix 10](#_Toc521415928)

[Sample Dataset 10](#_Toc521415929)

[Sample codes 11](#_Toc521415930)

[References 13](#_Toc521415931)

# Objective

## What are you setting out to prove or predict?

Insurance companies invest a lot of time and money to develop models that accurately forecast people’s medical expenses because they make profit when they collect more in yearly premiums than they spend on medical care to their beneficiaries. The goal of this analysis is to use the dataset to understand the relationship between explanatory variables and the response variable and build a model that predicts medical care expenses for such population segments.

## What is your rationale for there being a correlation in the data that you’re looking to confirm and/or exploit?

Although medical expenses are difficult to estimate, because certain conditions that are more costly are rare and seemingly random, some conditions are more prevalent for certain segments of the population. For instance, lung cancer is more likely among smokers than non-smokers, and heart diseases may be more likely among the obese. Thus, the greater the risk associated with certain health conditions, the more expensive one’s medical expenses will be.

# Data Preparation:

## What was your data source? How good was the data quality?

The dataset [“Medical Cost Personal Datasets”](https://www.kaggle.com/mirichoi0218/insurance) is a standard machine learning data set retrieved from Kaggle: an open platform for predictive modeling and analytics competition using real life datasets provided by companies and users. However, the original source of the dataset is from the book “Machine Learning with R” by Brett Lantz that introduces machine learning using R.

The overall quality of the data was very good as there were no missing values, unknown/inapplicable values. The full dataset consists of 1338 rows and 7 columns that represent information about person’s age, sex, BMI, number of children/dependents, smoking activity, region and charges (medical expenses). The data types vary from integer, float to object.

## What did you need to do to procure it? What tools or code did you need to use to prepare it for analysis?

The full dataset was procured by downloading the file (*insurance.csv*) [“Medical Cost Personal Datasets”](https://www.kaggle.com/mirichoi0218/insurance) from Kaggle. After importing the Pandas library to Jupyter Notebook, the full dataset was loaded into a DataFrame using the *pandas.read\_csv* method. Additionally, other libraries such as Scikit-Learn, Matplotlib, Seaborn and the StatsModels module was imported for the analysis. Since the data did not have any missing unknown and inapplicable values there was not much cleaning to do when preparing for analysis.

## What challenges did you face?

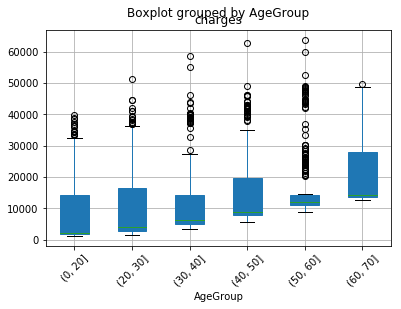
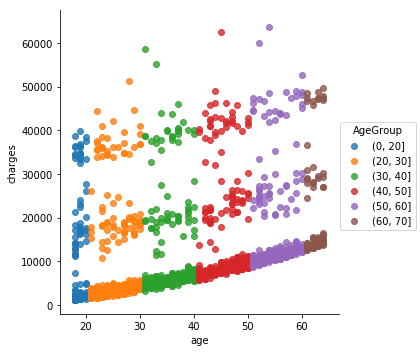
Among 7 columns in the dataset, ***sex****,* ***smoker*** and ***region*** variables were categorical variables that are non-numerical. Since regression models and machine learning models are based on mathematical functions, having categorical data may not yield the best performance. Thus, categorical variables that are not quantifiable needed to be converted to numerical values using the encoding process. There is a risk using the *LabelEncoder* class from Scikit-Learn library because the model could assume an ordinal relationship between categorical values and may result in less accurate performance. In this case, the *OneHotEncoder* class was applied where there is no relationship or order introduced between categories.

# Analysis

As mentioned above, the goal of this analysis is to use multiple regression, one response variable ***(charges)*** and many explanatory variables where they may be simultaneously connected, to estimate medical care expenses.

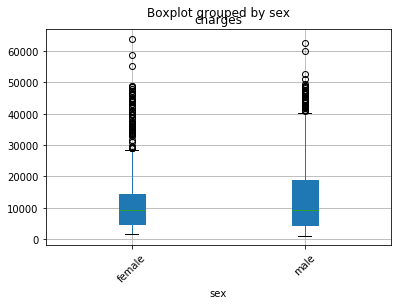
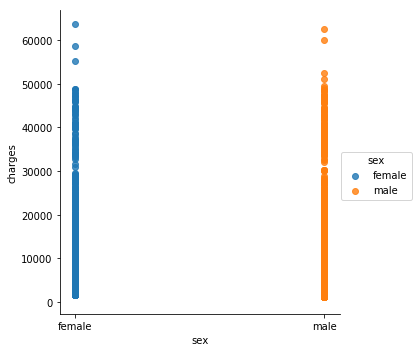
The following section will examine the relationship between each explanatory variables and the response variable to identify variables that may not be helpful in the model.

**Age vs charges**

****

The first explanatory variable to examine is the ***age*** variableand the values range from 18 to 64. ***age*** values were segmented into bins of 10 years to better identify any prevalent trends. The plot on the left indicates that although there is no clear linear relationship, in general, as age increases medical expenses increase as well. The minimum medical expenses for those in *AgeGroup* 60 to 70 is higher than that of those in *AgeGroup* 0 to 20. The boxplot on the right confirms these statements: the median medical expenses increase as age increases. Thus, ***age*** variable seems to have some relationship with ***charges***.

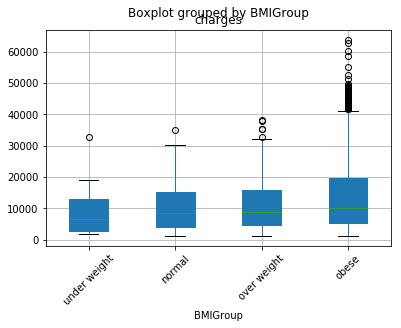
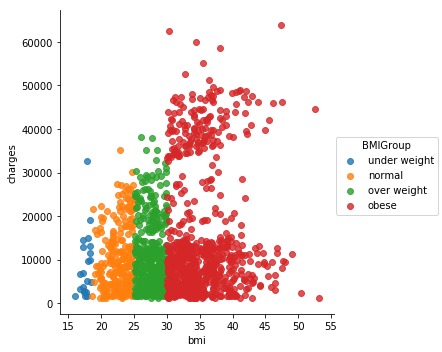
**Sex vs charges**

****

The ***sex*** variable is a two-level categorical variable indicating whether the person is a female or male. However, as mentioned above categorical variables were encoded using the *OneHotEncoder* class for better results (0=female, 1=male). By looking at the plot on the left, ***sex*** variable does not seem to greatly influence medical expenses. Although the interquartile range is greater for male than that of female, the median medical expenses seem very similar for female and male. Thus, ***sex*** variable does not seem to be helpful when added into the model.

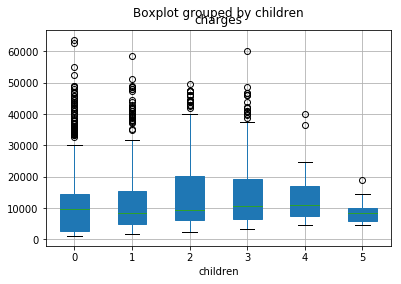
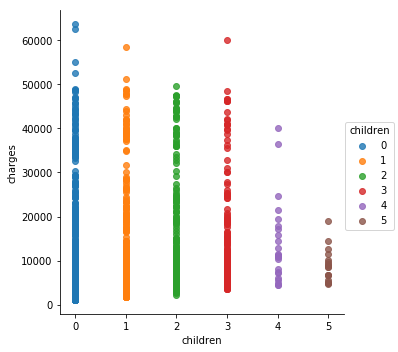
**BMI vs. charges**

Body mass index (BMI) is one of many factors that allows people to quickly assess one’s health risk based on one’s height and weight. Similar to ***age***values, ***bmi*** values were segmented into 4 groups representing whether a person is under weight (), normal over weight ( or obese to better identify any prevalent trends.

****

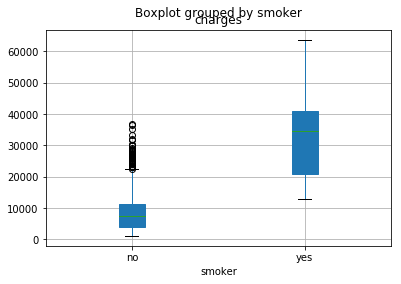
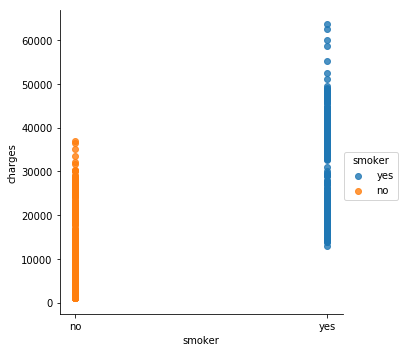
The plot on the left shows that majority of the people in the dataset were obese while the least number of people were underweight. The scatter plot as well as the box plot indicates that those who pay over $40,000 are the ones who are categorized as obese. Moreover, the statistics summary output showed that the mean medical expenses for obese people were $15,560.93 while for people in the normal group paid an average of $10,435.44. Obese people are at greater risk of developing conditions such as diabetes, heart disease and some types of cancer, which are also more costly to treat. Thus, ***bmi***variable seems to have some relationship with the ***charges***.

**Children vs. charges**



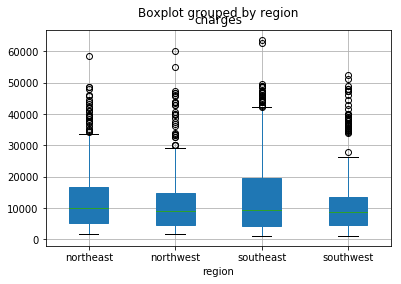
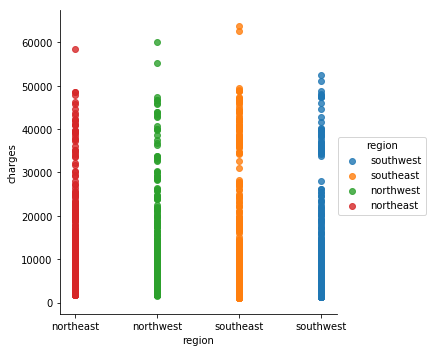
The **children** variable indicates how many children or number of dependents the person’s health insurance covers. The plot on the left shows that medical expenses are quite similar for the ***children*** value from 0 to 3. However, medical expenses decrease for those who have 4 and 5 children because according to [Blue Cross](https://www.bcbsm.com/index/health-insurance-help/faqs/topics/buying-insurance/family-size-impact-cost.html), when a person has more than 3 children under the age of 21, he or she only pays health plan for the three oldest when all members are covered. The ***children*** variable does not seem to be a very important explanatory variable.

**Smoker vs charges**

****

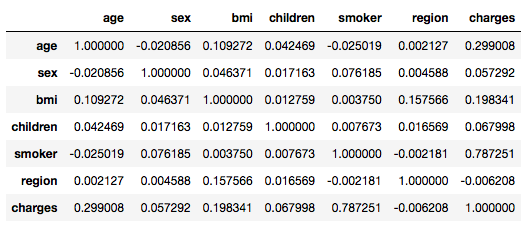
The ***smoker*** variable is a two-level categorical variable that indicates whether or not the person smokes. Similar to the ***sex*** variable, the values were encoded by using the *OneHotEncoder* class (0=non-smoker, 1=smoker). The plot on the left suggests that medical expenses for smokers are generally higher than non-smokers. Moreover, the boxplot shows that median medical expenses for smokers are around 4 times higher than non-smokers. Since ***smoker*** variable seemed to have a strong relationship with ***charges*** the single-variable model was created, and the regression output indicated that the value was 0.62. The coefficient for the linear model predicting medical expenses based on smoking activity was 23620, which means that the model predicts smokers to spend an extra $23,620 compared to those who don’t smoke. The regression output shows that the p-value for ***smoker*** variable is 0, indicating there is strong evidence that ***smoker*** variable is related to **charges**.

**Region vs charges**

****

The last explanatory variable to examine is the ***region*** variable. The plot on the left shows that there seems to be no significant difference in medical expenses among 4 regions and by looking at the boxplot, the minimum as well as the median medical expenses values are similar among 4 regions indicating that ***region***variable seems to be unhelpful when predicting medical expenses.

**Correlation between variables**

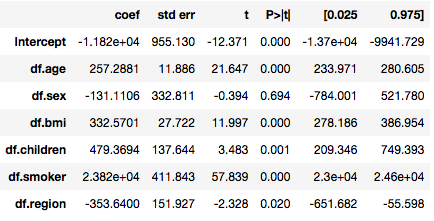


After examining the relationship between each explanatory variables and the ***charges***, ***smoker*** variable seemed to have the strongest correlation. This can be verified by using Pandas *df.corr()* method: ***smoker*** vs ***charges*** (0.787), followed by ***age*** vs ***charges*** (0.299) and ***bmi*** (0.198). Another thing to note is that predictor variables are not correlated among each other, so collinearity will not be an issue in this analysis.

**Building a model**

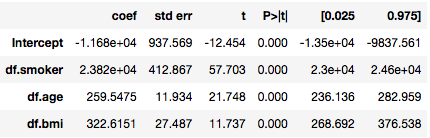
The goal is to fit a model that includes all potentially important variables simultaneously. This is helpful in evaluating the relationship between a predictor variable and the response variable while controlling for the potential influence of other variables.

Model 1: Full model



First model is a full model that includes all explanatory variables and has an adjusted value of 0.750. This means that 75.0% of the variation of charges could be explained by the set of explanatory variables that are included in the model. The p-value for the ***sex*** variable is greater than 0.05, which confirms the assumption earlier that the ***sex*** variable does not seem to be a significant variable predicting medical expenses. The ***region*** variable can also be excluded from the model as it doesn’t help increasing the accuracy of the model. Among many coefficients from the regression output, the coefficient for ***smoker*** variable is worth mentioning because it means that those who smoke are likely to spend $23,823 more than those who don't smoke. This solidifies the finding from earlier that ***smoker*** variable is strongly correlated with ***charges***.

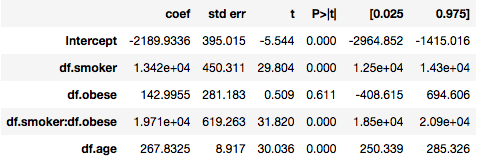
Model 2: Age & Smoker& Bmi



The second model includes ***smoker, age,*** and ***bmi*** variables since these 3 were identified as helpful variables to add in the model from individually examining the relationship between explanatory variables and the response variable, ***charges***. The adjusted value for this model was 0.747, which is slightly lower than the full model above.

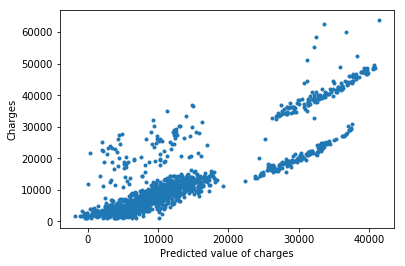
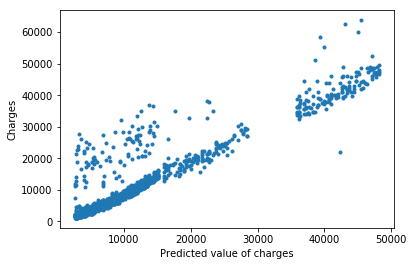
Since it was observed from the previous section that those who pay over $40,000 were the ones categorized as being obese, a new categorical variable ***obese*** was created to examine its relationship with medical expenses. Once again, *OneHotEncoder* class was used to convert the categorical values into numerical values (0=no, 1=yes).

Model 3: Derived variable

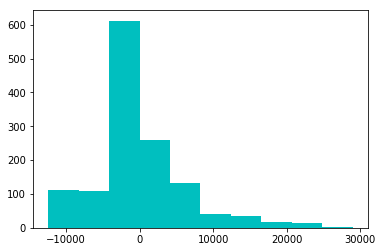
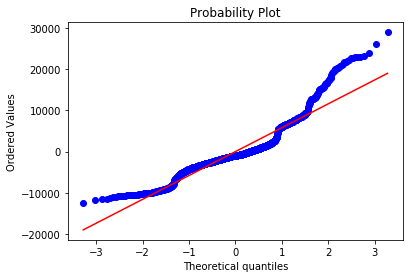


The new variable ***obese*** imposes a high penalty to those who smoke and are obese. If a person is a smoker, his or her medical expenses is $13,420 higher than those who don’t smoke. However, if the person smokes and is obese, he or she will likely spend $19,710 on medical expenses. By adding the derived variable, ***obese,*** the prediction of the model improved with an adjusted of 0.858 which implies 85.8% of variation of charges can be explained by the explanatory variables in the model.

**Real vs Prediction**

The scatterplot on the left shows the relationship between ***charges vs predicted value***using model 2. The scatterplot on the right shows the relationship between ***charges vs predicted value*** using model 3. The scatterplot using model 3 seems to have a more distinct linear trend compared to the one using model 2. Although accuracy of the model is important, it’s also important to build a model that is simple. Thus model 2 will be used for predicting medical expenses in this analysis.

****

The plot on the left is a normal probability plot which is helpful in identifying observations that might be outliers. By looking at the normal probability plot for residuals, the model seems to have quite a lot of residuals that appear to be outliers, and this indicates that there is a long tail in the distribution of residuals. This is confirmed by looking at the histogram of residuals on the right: the residuals are extremely right skewed.

Although the model isn’t quite the best model to predict medical expenses, the model can be tested to predict medical expenses of new data. Let’s assume that a person A is a smoker, 56 years old and has a bmi value of 26, the expected medical expenses for person A using model 2 is $35,069.50. On the other hand, let’s assume that person B is a non-smoker, 55 years old and has a bmi value of 23. The expected medical expenses for person B using model 2 is $10,018.43. This really assures how a person being a smoker could increase one’s medical expenses by almost 3 times. In order to test the difference in medical expenses when a person is obese or not, let’s assume person C is a non-smoker, 30 years old and has a bmi value of 33. The expected medical expenses for person C using model 2 is $ 6755.89. On the other hand, let’s assume that person D is a non-smoker, 32 years old and has a bmi value of 21. The expected medical expenses for person B using model 2 is $ 3403.61. This confirms that a person being obese definitely increases medical expenses.

# Conclusion

In this project, 6 explanatory variables were examined to determine their relationship and importance in building a model predicting the medical expenses, ***charges.*** As a result, variables that were most likely to affect a person’s medical expenses was the ***smoker*** variable, followed by ***age*** and ***bmi*** variables. When a single-variable model for ***charges vs. smoker*** was created, the regression output result indicated that the coefficient of ***smoker*** variable was 23620, meaning that the model predicts smokers to spend an extra $23,620 compared to those who don’t smoke. Although there was no clear linear relationship between ***charges vs age***, medical expenses generally increased as age increased. The full model that included all explanatory variables had an adjusted adjusted value of 0.750, meaning that 75.0% of the variation of charges could be explained by the set of explanatory variables that are included in the model. The second model included 3 explanatory variables (***smoker, age, bmi***) that were most relevant to predicting medical expenses and the adjusted value was 0.747, slightly lower than model 1. Since we observed that majority of the people in the dataset were obese and those who spent over $40,000 were the ones categorized as obese, a new categorical variable ***‘obese’*** was derived from the ***bmi*** variable. The purpose was to examine if a person being obese actually influenced medical expenses while still including the ***age, smoker*** variables. By using the formula , we significantly improved the adjusted of the model to 0.858. The important finding using this model was that if a person is a smoker, his or her medical expenses is $13,420 higher than those who don’t smoke. However, if the person smokes and is obese, he or she will likely spend $19,710 (additional $6,000) on medical expenses. When insurance companies are developing models that accurately forecast people’s medical expenses they should use model 3 as their base model and further engineer the model to improve its prediction accuracy. In order to keep the model simpler than real life, model 2 was chosen in this project to predict medical expenses of new test data. Although diagnostic plots above show that our model does not fully meet the four assumptions of a multiple regression model, the analysis still confirmed that people who smoke, are obese and older have higher medical expenses and health insurance companies should charge higher premium on these people.

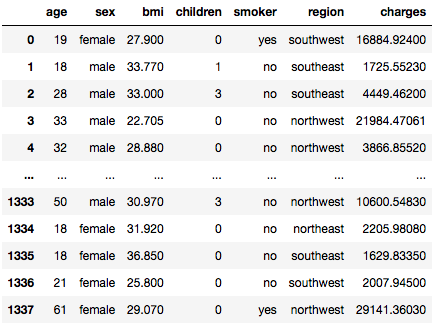
# Appendix

## Sample Dataset

* Data dictionary

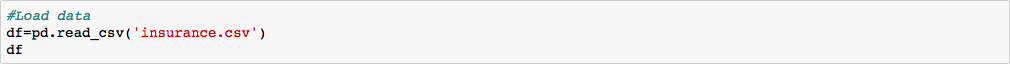
|  |  |
| --- | --- |
| age | Age of person |
| sex | Gender of person: Female vs Male |
| bmi | Body Mass Index: body fat based on height and weight |
| children | Number of children/dependents |
| smoker | Smoking activity: Yes vs No |
| region | Residential location of the person in the US: Northeast, Southeast, Southwest, Northwest |
| charges | Medical expenses |

* Sample Data

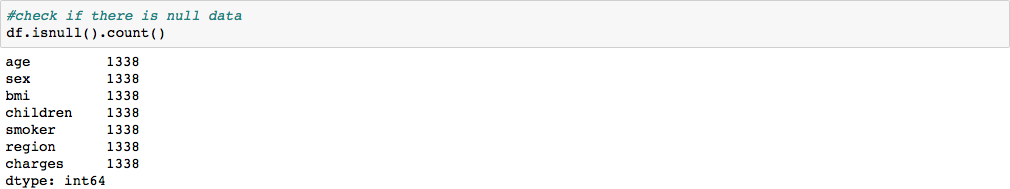


## Sample codes

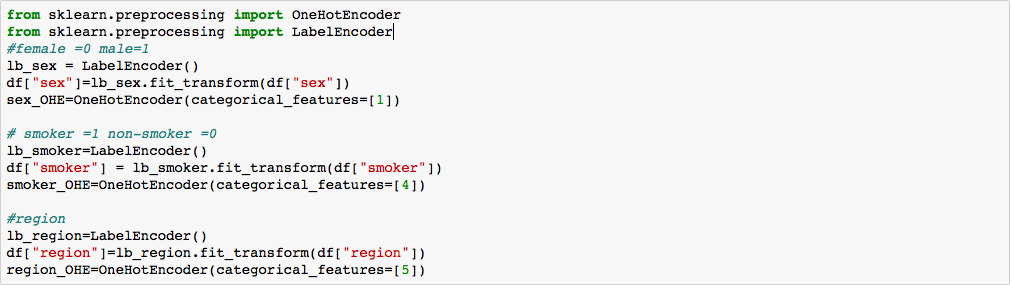
* Load Dataset



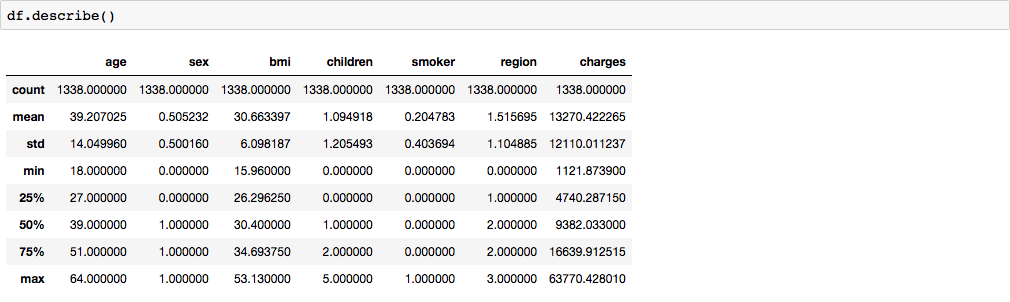
* Checking missing values



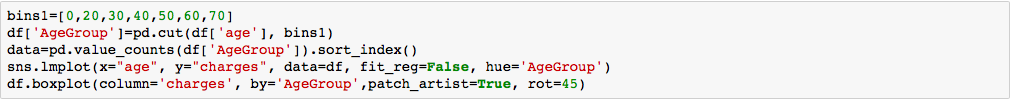
* Encoding categorical variables using OneHotEncoder



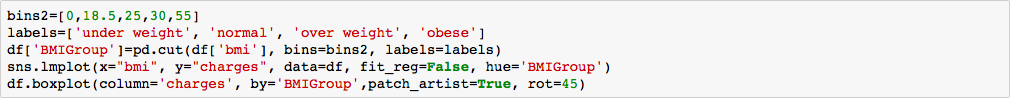
* Generate descriptive statistics



* Group ***age*** values into bins



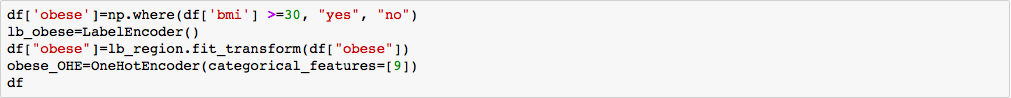
* Group ***bmi*** values into underweight, normal, overweight, obese



* Visualizing ***explanatory variable vs charges*** sample code



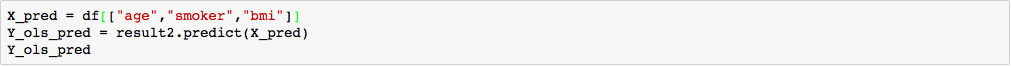
* Create ***obese*** column & encode using OneHotEncoder



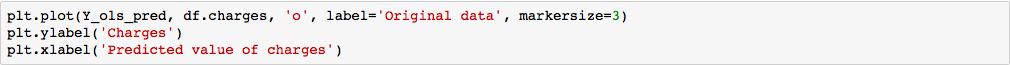
* Ordinary Least Square regression (***obese \* smoker***)



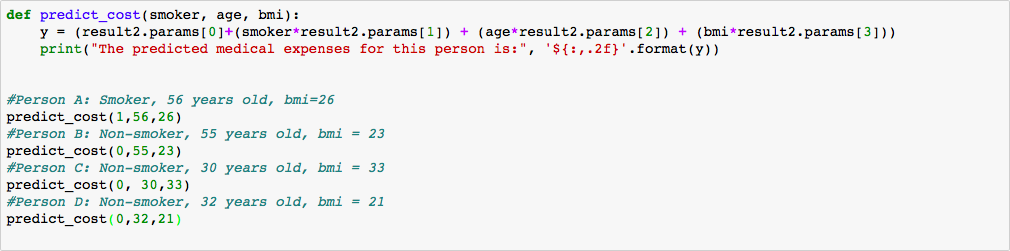
* Predict medical expenses (charges) using the model



* Plot real values vs predicted values



* Predict medical expenses by inputting new data



# References

* Kaggle: “Medical Cost Personal Datasets”

<https://www.kaggle.com/mirichoi0218/insurance>

* BlueCross

<https://www.bcbsm.com/index/health-insurance-help/faqs/topics/buying-insurance/family-size-impact-cost.html>