**MEDICAL INSURANCE PREMIUM PREDICTION**

# CONTENTS

* [Executive summary](#_TOC_250003)
* Problem statement
* Background
* Methodology
* Data Source
* Introduction
* Dependent and independent variables
* Descriptive statistics
* Scatter plots
* Correlation matrix and multicollinearity
* Omitted variable bias
* Summary of linear models
* Conclusion
* Python code

# INTRODUCTION:

The Insurance Premium is influenced by various factors. It is complex function of various factors namely Sex, Region, Body mass index, Children and Smoking habits. The historical data set of medical insurance are collected from various regions of US.

The data consists of medical expenses paid by people under certain conditions characterized by attributes such as location, gender, number of children covered by health insurance / Number of dependents, Age of primary beneficiary. We have analyzed the effect of each of these factors on the medical insurance premium and have tried to construct various regression models based on these factors and included factors that played a significant role.

# VARIABLES:

**Dependent variable (Y):**

Medical insurance premium:Individual medical costs billed by health insurance.

**Independent variables (X):**

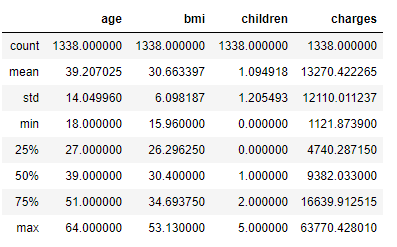
Following factors affect the premium of medical insurance:

* **Body mass index**: BMI is an indicator of the fitness level of insurance bearer. Disproportionate BMI is a sign of an unhealthy body likely to get affected by diseases. So the premium charged from such people are likely to be higher.
* **Age**: The age group given in data is 18-64. With increasing age people are more prone to different ailments which should get reflected in the premium costs as well.
* **Sex**: This binary attribute provide details about the gender of the primary beneficiary
* **Smoker**: This is also a binary attribute which can have effect on the health of the insurance bearer. Risk of getting diseases for a smoker is much more than a nonsmoker. So, the nonsmokers will have probably less insurance amount compared to smokers.
* **Region**: The beneficiary's residential area in the US, northeast, southeast, southwest, northwest. Health of people is affected by the demographic factors and the socio economic development of that area. So region turns out to be an important factor influencing the premium amount.
* **Children**: Number of children covered by health insurance / Number of dependents. The health insurance covers the individual as well as his/her dependents. So more the number of children a person has more will be the number of people covered which will increase the premium costs.

# DESCRIPTIVE STATISTICS:

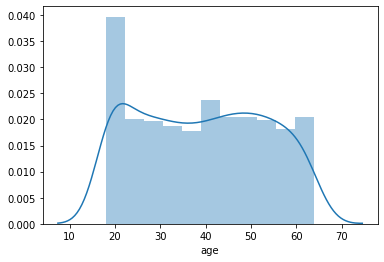
Descriptive statistics provide simple summaries about the sample and about the observations that have been made. Such summaries may be either quantitative, i.e. summary statistics, or visual, i.e. simple-to-understand graphs. These summaries form the basis of the initial description of the Medical insurance premium prediction data here.

Some measures that are commonly used to describe a data set are measures of central tendency and measures of variability or dispersion. Measures of central tendency include the mean, median and mode, while measures of variability include the standard deviation (or variance), the minimum and maximum values of the variables, kurtosis , quartiles and skewness.

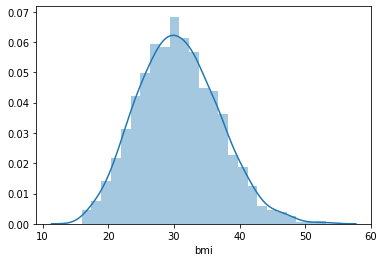
****

**DISTRIBUTIONS OF CONTINUOS INDEPENDENT VARIABLES:**

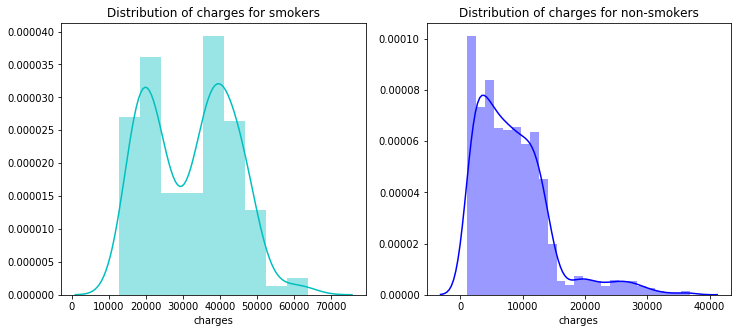
**DISTRIBUTION OF AGE:**

****

**DISTRIBUTION OF BMI:**

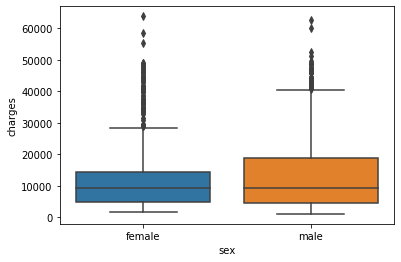
****

**DISTRIBUTION OF SMOKERS:**

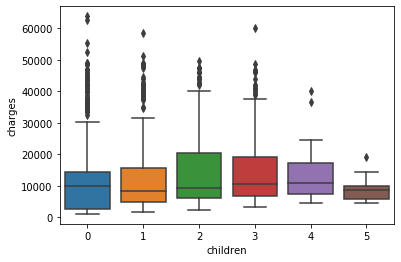
****

**BOX PLOT FOR INDEPENDENT CATEGORICAL VARIABLES:**

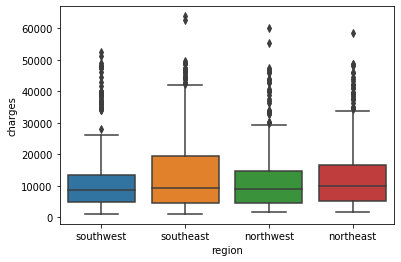
**BOX PLOT FOR GENDER ATTRIBUTE**

****

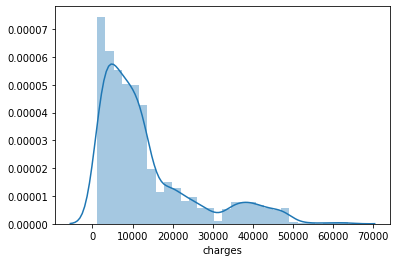
**BOX PLOT FOR NUMBER OF CHILDREN ATTRIBUTE**

****

**BOX PLOT FOR REGION ATTRIBUTE**

****

**DISTRIBUTIONS OF DEPENDENT VARIABLE (MEDICAL CHARGES):**

****

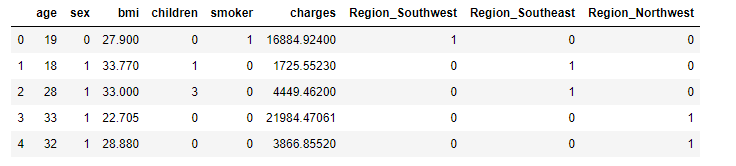
# DUMMY VARIABLE TRAP:

Dummy variables are required to represent a categorical variables the number of dummy variables depends on the number of values that particular categorical variable can take. If we have to represent a categorical variable that can take N different values, we need to define N - 1 dummy variables.

Here in this dataset we have an attribute Region which can have four categories Southwest, Northwest, Southeast, Northeast, In order to include this attribute in the our regression model we need to encode this to a numerical values. While converting this categorical attribute to numerical attribute we need to take care of dummy variable trap issue.

When dummy variables are defined, we need to be careful or else we might end up defining too many variables. If a particular categorical variable takes on N values, it is highly possible that we may define N dummy variables. If we define N dummy variables then we will end up in this trap called Dummy variable trap. This could lead us to linear dependence between these variables so you only need N - 1 dummy variables.

A Nth dummy variable is redundant as it carries no new information. And it creates a severe [multicollinearity](https://stattrek.com/statistics/dictionary.aspx?definition=multicollinearity) problem for the Regression analysis. In this dataset we have removed one region from the four available regions to overcome this problem. Relevant attributes after conversion is shown below.

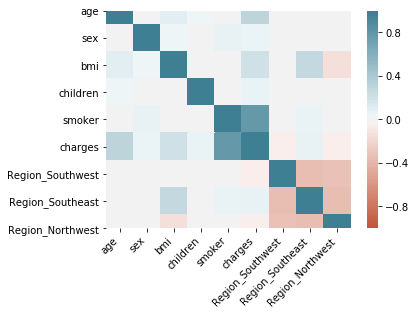


CORRELATION MATRIX AND MULTICOLLINEARITY

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two variables or bivariate data. In the broadest sense correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related.

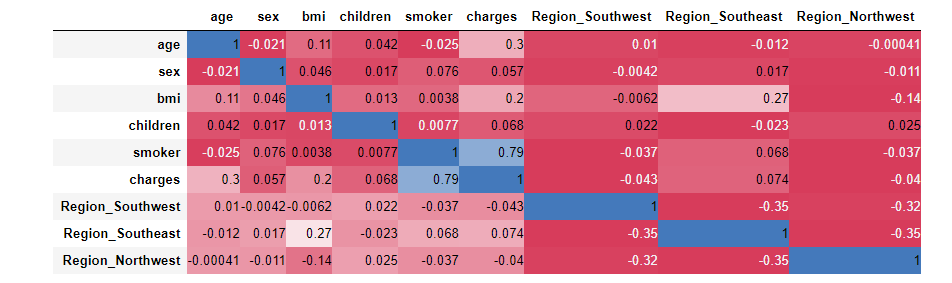
In statistical modelling, correlation matrices represents the relationships between variables. Also we need to be aware that correlation does not imply causation. In our problem of predicting medical expenses we are making use of the correlation matrix to identify the independent variable which has

Close relationship with the dependent variable moreover this correlation matrix also help us to identify the multicollinearity associated with the independent variables themselves. The correlation matrix for our data is depicted below.

****

**CORRELATION BETWEEN INDEPENDENT AND DEPENDENT VARIABLE:**

|  |
| --- |
| Region\_Southwest -0.043210 |
| Region\_Northwest -0.039905 |
| sex 0.057292 |
| children 0.067998 |
| Region\_Southeast 0.073982 |
| bmi 0.198341 |
| age 0.299008 |
| smoker 0.787251 |
| charges 1.000000 |



**MULTICOLLINEARITY FINDINGS:**

Multicollinearity is a concept in which one independent variable in a multivariate regression model can be linearly predicted from the other independent variables with a good degree of accuracy. In such a situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data.

In our data, we can observe a small correlation between the independent variables for Region southwest and the sex close 0.27 which is not substantial so we can confirm that no multicollinearity exists in our variables.

**MOST SIGNIFICANT INDEPENDENT VARIABLES FROM CORRELATION MATRIX:**

|  |
| --- |
| BMI 0.198341 |
| AGE 0.299008 |
| SMOKER 0.787251 |

# OMITTED VARIABLE BIAS:

Omitted variable bias occurs when two conditions are true-

* When the omitted variable is correlated with the included independent variable and
* When the omitted variable is a determinant of the dependent variable.

The omitted variables for the given problem statement are as follows-

* Eating habits
* Daily physical activity
* Drinking habits
* Genetics
* Quality of treatment
* Doctor qualification
* Frequency of smoking
* Brand of cigarette
* Years of smoking

SUMMARY OF LINEAR MODELS

Linear regression is a method in which we fit regression line between a dependent variable and one or more independent variables. We formulated 12 different models to understand the relationship of Medical expenses with other factors like BMI, Age etc. The list of models formulated in this study are listed below:

**Model 1:**

**Charges = 257.7226 \* Age + 3165.8850**

**(Slope) (Intercept)**

## 

## Std. error=937.149, t statistic=3.378 , P>|t|=0.001(for intercept);

|  |  |
| --- | --- |
|  |  |

## 95 percent Confidence interval for intercept (1327.440, 5004.330)

## 95 percent Confidence interval for slope (213, 301)

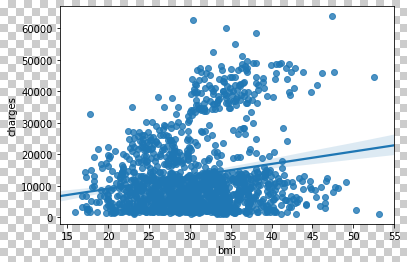
## R square =0.089

## As both the confidence intervals obtained after t test for intercept and slope does not contain zero we reject the null hypothesis (Reduced model explains the variance better). And hence we conclude that there is a relationship between age and charges. But the R square obtained is very less which signifies very less variance in charges is explained by age alone.

**Model 2:**

**Charges = 393.873 \* Bmi + 1192.9372**

**(Slope) (Intercept)**

****

## 

## Std. error=1664, t statistic=0.717, P>|t|=0.474(for intercept);

|  |  |
| --- | --- |
|  |  |

## 95 percent Confidence interval for intercept (-2027.974, 4458.849)

## 95 percent Confidence interval for slope (289.409, 498.337)

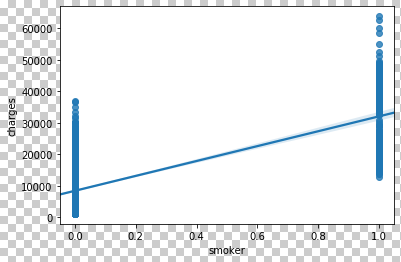
## R square =0.039

## The confidence intervals obtained after t test for intercept contain zero thus the intercept obtained is not significant so we do not reject the null hypothesis for intercept (intercept=0). The confidence interval for slope parameter does not contain 0. Which implies the significance of slope parameter. But the R square obtained is very less which signifies very less variance in charges is explained by body mass index alone.

**Model 3:**

**Charges = 2.362e+04 \* Smoker + 8434.2683**

**(Slope) (Intercept**)



## Std. error = 229.014, t statistic = 36.829, P>|t| = 0(for intercept);

## 95 percent Confidence interval for intercept (7985.002, 8883.535)

## 95 percent Confidence interval for slope (2.26e+04, 2.46e+04)

## R square =0.620

## As both the confidence intervals obtained after t test for intercept and slope does not contain zero we reject the null hypothesis (Reduced model explains the variance better).The p value is close to zero which signifies a strong relationship between smokers and charges. The R square obtained is 0.62 which signifies that about 60 % of the variance in premium charges are explained alone by the smoker or not binary attribute.

|  |  |
| --- | --- |
|  |  |

**Model 4:**

**Charges = 241.9308 \* Age + 332.9651 \* Bmi - 6424.8046**

**(X1) (X2) (Intercept)**

**Standard errors: X1= 22.298 X2 = 51.374 Intercept=1744.091**

**T statistic: X1= 10.850 X2 = 6.481 Intercept=-3.684**

**Confidence interval: X1= (198.187 285.674)**

**X2= (232.182 433.748)**

**Constant= (-9846.262 -3003.347)**

**R Squared= 0.117**

**Adjusted R Squared=0.116**

## From this model we came to know that there is a linear relationship between age, bmi and charges. With all |t| values greater than 1.96 we conclude that all the coefficients are statistically significant. No confidence intervals for coefficients of x1, x2 and intercepts contains 0. Thus we reject the null hypothesis that there is no relationship between the selected independent variables and charges.

## The F statistic obtained is very small than the significance level. For the model with no independent variables, the intercept-only model, all of the model’s predictions equal the mean of the dependent variable. Consequently, if the overall F-test is statistically significant, your model’s predictions are an improvement over using the mean.

**Model 5:**

**Charges = 274.87121\*Age + 2.386e+04\* smoker- 2391.62**

**(Slope) (Slope) (Intercept)**

**Standard errors: X1=12.544 X2=433.488 Intercept=528.302**

**T statistic: X1=22.069 X2=55.031 Intercept=-4.527**

**Confidence interval: X1 = (250.437 299.305)**

**X2 = (2.3e+04 2.47e+04)**

**Constant = (-3428.019 -1355.234)**

**R Squared= 0.721**

**Adjusted R Squared=0.721**

## From this model we come to know that there is a linear relationship between age, smoker and charges. With all |t| values greater than 1.96 we conclude that all the coefficients are statistically significant. No confidence intervals for coefficients of x1, x2 and intercepts contains 0. Thus we reject the null hypothesis that there is no relationship between the selected independent variables and charges.

## The F statistic obtained is very small than the significance level. For the model with no independent variables, the intercept-only model, all of the model’s predictions equal the mean of the dependent variable. Consequently, if the overall F-test is statistically significant, your model’s predictions are an improvement over using the mean. The R square obtained is 0.72 which is relatively better as compared to all other models. Around 70 percent of the variance is explained by these two dependent variables (ie Bmi and age).

|  |  |
| --- | --- |
|  |  |

**Model 6:**

**Charges = 388.0152\*Bmi + 2.359e+04\*smoker- 3459.0955**

**(Slope) (Slope) (Intercept**)

**Standard errors: X1=31.787 X2=480.18 Intercept=998.279**

**T statistic: X1=-3.465 X2=12.207 Intercept=49.136**

**Confidence interval: X1= (325.656 450.374)**

**X2= (2.7e+04 2.47e+04)**

**Constant= (-5417.463 -1500.728)**

**R Squared= 0.658**

**Adjusted R Squared=0.657**

## From this model we come to know that there is a linear relationship between Age, Smoker and charges. With all |t| values greater than 1.96 we conclude that all the coefficients are statistically significant. No confidence intervals for coefficients of x1, x2 and intercepts contains 0. Thus we reject the null hypothesis that there is no relationship between the selected independent variables and charges.

## The F statistic obtained is very small than the significance level. For the model with no independent variables, the intercept-only model, all of the model’s predictions equal the mean of the dependent variable. Consequently, if the overall F-test is statistically significant, your model’s predictions are an improvement over using the mean. The R square obtained is 0.657 which is relatively better as compared to all other models. Around 65 percent of the variance is explained by these two dependent variables (ie BMI and smoker).

**Model 7:**

**Charges = 259.5475 \* Age+ 322.6157\* Bmi + 2.382e+ 04\*smoker - 2.382e+04**

**(Slope) (Slope) (Slope) (Intercept)**

**Standard errors: X1=259.5475 X2=322.6151 X3=2.382e+04 intercept=-1.168e+04**

**T statistic: X1=21.748 X2=11.737 X3=57.703 intercept=-12.454**

**Confidence interval: X1 = (236.136 282)**

**X2 = (268.692 376.538)**

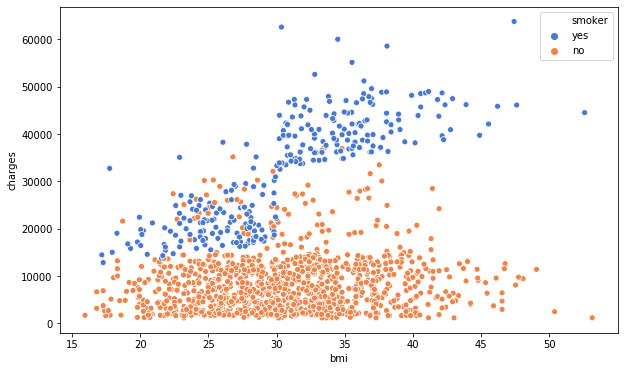
**X3 = (2.3e+04 2.46e+04)**

**R Squared= 0.747**

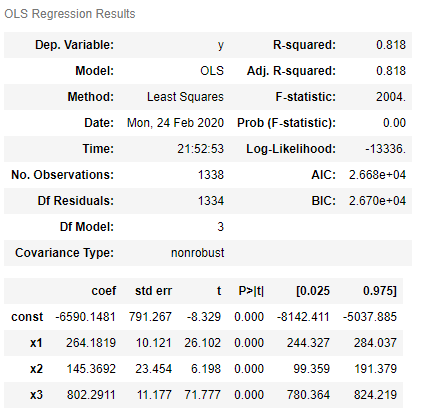
**Adjusted R Squared=0.747**

## This regression model has been applied with three independent variables (age, bmi, smoker) .The F statistic is 1316 and p value is close to 0. H0:   β1 = β2 = β3 = 0  H1:   βj ≠ 0, for at least one value of j. Thus we can safely reject the null hypothesis that all the coefficients are zero simultaneously. T statistics obtained are all much higher than 1.96 which proves their statistical significance. No confidence interval contains 0 in it. Hence we reject the null hypothesis about all the coefficients individually. This model has the highest R squared value. 75 percent of the variance is explained by these 3 attributes

**MODEL 8 (With Additional Attribute):**

The previous models were based on the attributes that were readily available in the dataset. By making use of the available features we achieved an R – square value of about 0.75. Moreover, we can create one more feature BMI OF SMOKERS based on the two available variables smoker (binary) and body mass index (continuous) and we can incorporate this BMI OF SMOKERS attribute in our model to see a significant improvement in the R- Square value. The details of the same is shown in the below model.

The summary of the additional model with inclusion of this new attribute is described below.



**MODEL-8**

**Model 8:**

**Charges = 264.18 \* Age + 145.37 \* bmi + 802.29 \* bmi \* smoker- 6590.15**

**(Slope) (Slope) (Slope) (Intercept**)

**Standard errors: X1 = 10.121 X2 = 23.454 X3 = 11.177 Intercept = 791.267**

**T statistic: X1 = 26.102 X2 = 6.198 X3 = 71.777 Intercept= -8.329**

**Confidence interval: X1= (244.327 , 284.037)**

**X2= (99.36, 191.379)**

**X3= (780.36, 824.219)**

**Constant= (-8142.411 -5037.885)**

**R Squared= 0.818**

**Adjusted R Squared=0.818**

From this model we come to know that there is a linear relationship between Age, BMI, BMI of smoker and charges. With all |t| values greater than 1.96 we conclude that all the coefficients are statistically significant. No confidence intervals for coefficients of x1, x2 and intercepts contains 0. Thus we reject the null hypothesis that there is no relationship between the selected independent variables and charges.

The R square obtained is 0.818 which is very much better as compared to all other models. Around 82 percent of the total variance is explained by these independent variables (ie Age, BMI, and BMI of smoker).

# CONCLUSION

* After trying to fit a linear model for all the possible combinations of independent variables we can conclude that the multiple linear regression model with three independent variables viz age, bmi and bmi of smokers explains the variance in dependent variable the best.
* There was an increasing trend in the R square values obtained when we went on increasing the number of independent variables.
* R squared value for the model is 0.818 which signifies that around 82 percent of the variance has been explained by these variables
* After analysis of model we found that there can be omitted variable bias due to various factors like

1. No records of preexisting medical conditions

2. Profession of individual: High risk profession are more likely to be charged higher premiums

3. Family medical history

# APPENDIX

**PYTHON CODE**

**IMPORTING LIBRARIES**

**import** **numpy** **as** **np**

**import** **pandas** **as** **pd**

**import** **matplotlib.pyplot** **as** **plt**

**import** **seaborn** **as** **sns**

**IMPORTING THE DATASET**

dataset = pd.read\_csv('insurance.csv')

In [3]:

dataset.head()

**IDENTIFYING MISSING VALUES**

dataset.isnull().sum()

**CONVERTING CATEGORICAL ATTRIBUTES TO NUMERICAL**

**from** **sklearn.preprocessing** **import** LabelEncoder

*#sex*

labelencoder = LabelEncoder()

labelencoder.fit(dataset.sex.drop\_duplicates())

dataset.sex = labelencoder.transform(dataset.sex)

*# smoker or not*

labelencoder.fit(dataset.smoker.drop\_duplicates())

dataset.smoker = labelencoder.transform(dataset.smoker)

dataset.head()

dataset['Region\_Southwest'] = dataset.region.map({'southwest':1,'southeast':0,'northwest':0,'northeast':0})

dataset['Region\_Southeast'] = dataset.region.map({'southwest':0,'southeast':1,'northwest':0,'northeast':0})

dataset['Region\_Northwest'] = dataset.region.map({'southwest':0,'southeast':0,'northwest':1,'northeast':0})

dataset['Region\_Northeast'] = dataset.region.map({'southwest':0,'southeast':0,'northwest':0,'northeast':1})

dataset.head()

**AVOIDING DUMMY VARIABLE TRAP**

dataset.drop('Region\_Northeast',axis=1,inplace = **True**)

dataset.drop('region',axis=1,inplace = **True**

**CHECKING CORRELATION BETWEEN VARIABLES**

dataset.corr()['charges'].sort\_values()

corr = dataset.corr()

ax = sns.heatmap(

corr,

vmin=-1, vmax=1, center=0,

cmap=sns.diverging\_palette(20, 220, n=200),

square=**True**

)

ax.set\_xticklabels(

ax.get\_xticklabels(),

rotation=45,

horizontalalignment='right'

)

cmap = cmap=sns.diverging\_palette(5, 250, as\_cmap=**True**)

**def** magnify():

**return** [dict(selector="th",

props=[("font-size", "10pt")]),

dict(selector="td",

props=[('padding', "0em 0em")]),

dict(selector="th:hover",

props=[("font-size", "12pt")]),

dict(selector="tr:hover td:hover",

props=[('max-width', '400px'),

('font-size', '12pt')])

]

corr.style.background\_gradient(cmap, axis=1)\

.set\_properties(\*\*{'max-width': '100px', 'font-size': '10pt'})\

.set\_caption("Hover to magify")\

.set\_precision(2)\

.set\_table\_styles(magnify())

g = sns.pairplot(dataset,kind="reg")

ax = sns.boxplot(x="sex", y="charges", data=dataset)

ax = sns.boxplot(x="children", y="charges", data=dataset)

ax = sns.boxplot(x="region", y="charges", data=dataset)

sns.regplot(x="age", y="charges", data=dataset)

sns.regplot(x="bmi", y="charges", data=dataset)

plt.figure(figsize=(10,6))

ax = sns.scatterplot(x='bmi',y='charges',data=dataset,palette='muted',hue='smoker')

<https://stattrek.com/multiple-regression/dummy-variables.aspx>

<https://en.wikipedia.org/wiki/Correlation_and_dependence#Correlation_matrices>

<https://en.wikipedia.org/wiki/Descriptive_statistics>

<https://en.wikipedia.org/wiki/Multicollinearity>