**Project report**

**on**

**Health Insurance claim**

Submitted by:

**Eswar Katakam**

GLCA March-23

Contents

[a) Identify the categorical and continuous variables 3](#_Toc135746440)

[b) Make Histograms and box plots (univariate analysis) for continuous variables and do a correlation analysis (multivariate analysis) 3](#_Toc135746441)

[c) Make relevant Pivot tables and charts for: 3](#_Toc135746442)

[d) Region-wise smokers vs Non-smokers analysis with one or more pivot table and charts 5](#_Toc135746443)

[e) Region-wise charges for smokers vs non-smokers 5](#_Toc135746444)

[f) Has charges got something to do with the number of dependents ? 6](#_Toc135746445)

[g) Do a similar dependants-charges analysis, Region-wise 6](#_Toc135746446)

[h) Do at least one more pivot table and chart of your own choice on the remaining variables 6](#_Toc135746447)

[i) Give your understanding from the patterns observed in point (b) 7](#_Toc135746448)

[j) Give your interpretation for observations made in point (c) 7](#_Toc135746449)

[2. Edit the data as following, to obtain dummy variables: 7](#_Toc135746450)

[a) Sex : Replace all the “Males” with “1” and “Females” with “0”, 7](#_Toc135746451)

[b) Smoker: Replace all the “Smokers” with “1” and “Non-smokers” with “0”. 7](#_Toc135746452)

[c) Region: We always create one less category column for the dummy data w.r.t 7](#_Toc135746453)

[3. Do a descriptive summary analysis for the edited data. 8](#_Toc135746454)

1. Perform the Exploratory Data Analysis on the data.

## a) Identify the categorical and continuous variables

Categorical Variables:

Smoker,Region,Children,Sex

Continuos Variables:

Age, BMI,Charges

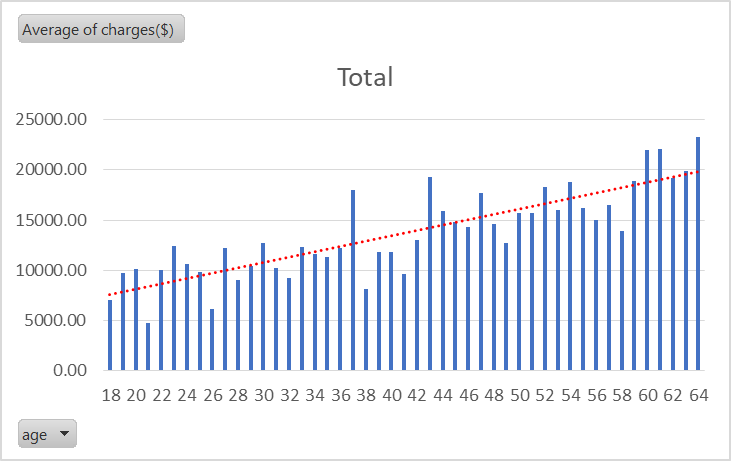
### b) Make Histograms and box plots (univariate analysis) for continuous variables and do a correlation analysis (multivariate analysis)

### c) Make relevant Pivot tables and charts for:

i) Male/Female ratio and share information on which gender has more smokers

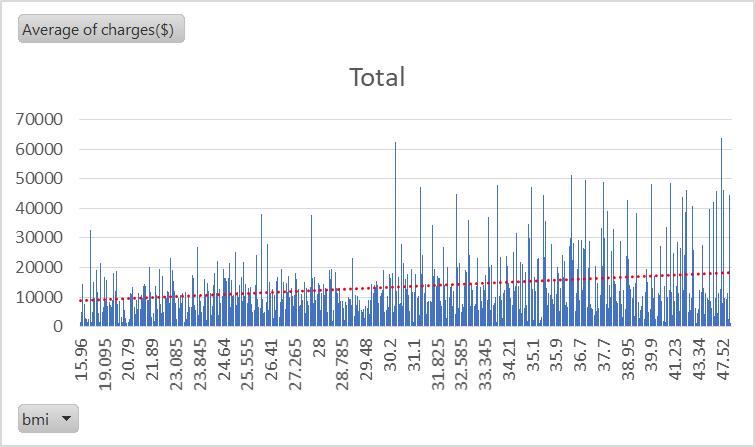
From the above Pie-Chart, we can say Male’s has **58%** Smokers among Female’s has **42%.**

ii) Charges vs Age

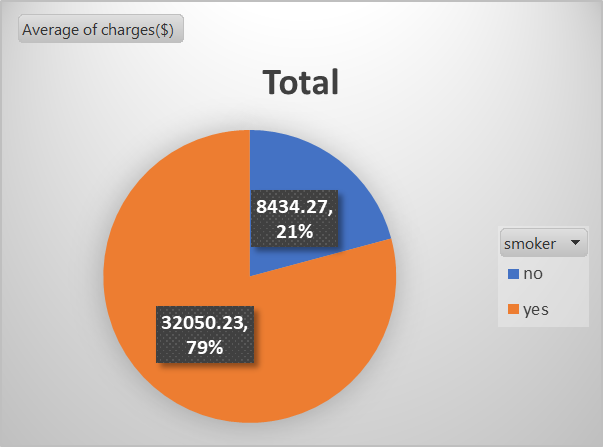


From the above Bar Graph, Along with the age Charges per individual also Increases. Means depending on Life Expentency the Insurance policy has more charges while increasing in Age might be depending on health problems.

iii) Charges vs BMI

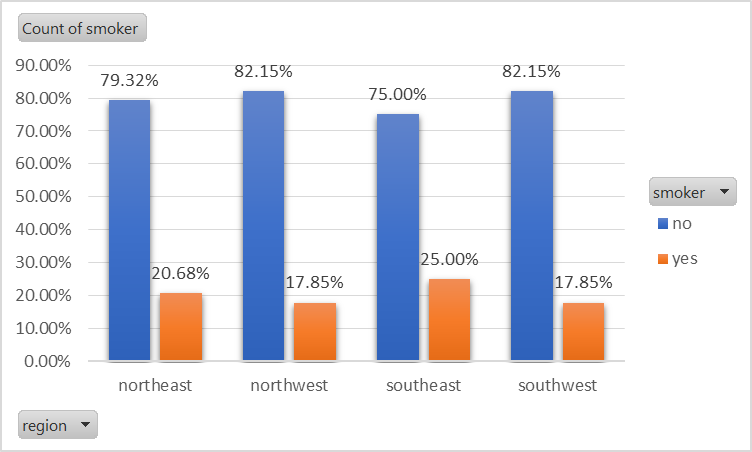


iv) Charges for Smokers vs Non-smokers



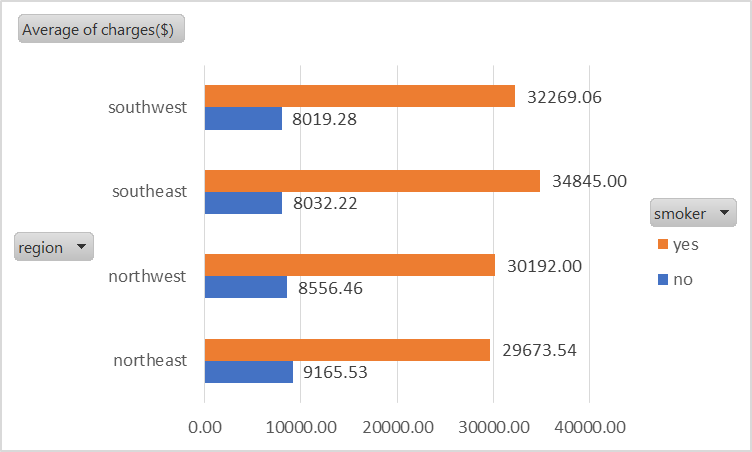
Charges for smokers are Higher than the average charges for Non-smokers.

## d) Region-wise smokers vs Non-smokers analysis with one or more pivot table and charts



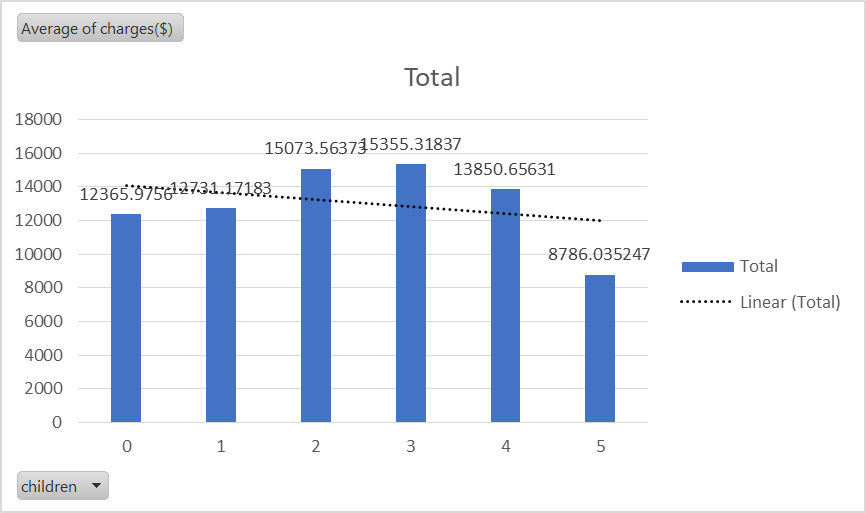
All the regions have more than **75%** and average of **79.6%** of non-smokers and **southeast** region has the more percentage of smokers among all above regions.

## e) Region-wise charges for smokers vs non-smokers

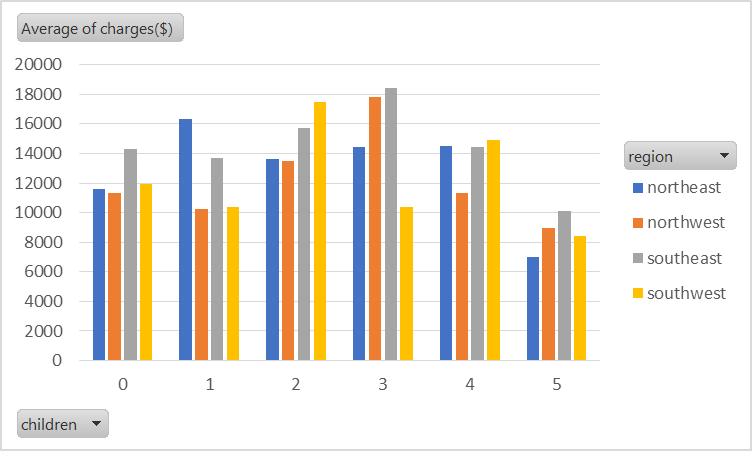


From the above chart, we can say that southeast region has the more number of smokers and Average charges also higher, similarly the lowest number of smokers having least average charges.Northeast region has highest average charges for non-smokers.

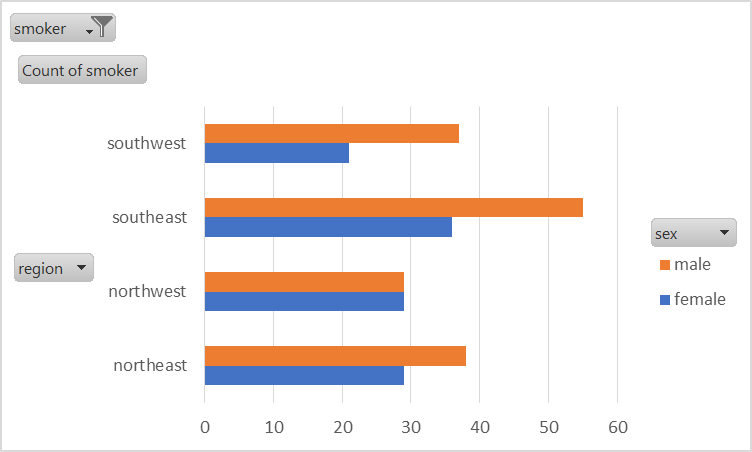
## f) Has charges got something to do with the number of dependents ?



## g) Do a similar dependants-charges analysis, Region-wise



## h) Do at least one more pivot table and chart of your own choice on the remaining variables



## i) Give your understanding from the patterns observed in point (b)

## 

## j) Give your interpretation for observations made in point (c)

From 1) Pie-Chart, we can say Male’s has **58%** Smokers among Female’s has **42%**.

From 2) Bar Graph, Along with the age Charges per individual also Increases. Means depending on Life Expectancy the Insurance policy has more charges while increasing in Age might be depending on health problems.

From 3) bar graph, we can say that Average charges are increases along with **bmi .**

From 4) Charges for smokers are Higher than the average charges for Non-smokers.

# 2. Edit the data as following, to obtain dummy variables:

a) Sex : Replace all the “Males” with “1” and “Females” with “0”, creating numerical entries for gender this way will help you do analysis further. You can use the “Replace with Match entire cell content” option. Do a replace all to save time.

### b) Smoker: Replace all the “Smokers” with “1” and “Non-smokers” with “0”.

c) Region: We always create one less category column for the dummy data w.r.t the categories available for that original variable. So for Region, we will create three dummy columns, assuming “Northeast” as zero and omit the column for it. Now create three columns for “northwest”, “Southeast”, “Southwest”. Whichever row has “northwest” region as an entry will take “1” as an entry otherwise “0” in “northwest” column. Similarly in the “Southeast” column, whichever row had “southeast” as an entry will take “1” as the new entry and “0” for the rest of the column (Southeast). Do a similar operation on the “Southwest” column. Please refer to the below image for your understanding.

{ 2b }{ 2c }{ 2a }



3. Do a descriptive summary analysis for the edited data. Perform a Multiple Linear Regression analysis to identify which variables decide the insurance charges/billed insurance claim. Give your interpretation for the above analysis, do another set of regression analysis by dropping insignificant variables, if needed.







From descriptive analysis , we can consider following points:

Mean age for dataset given is 39.2 ,age range from 18 – 64 most of the people are having Age 18.

Average charges for the given data are 13270.4.

Most people are paid with charge 1639.56

And minimum charge is 1121.87 and maximum charge is 63770.

Here the model is built with only significant variables which are age , BMI, Children and smokers.

**1st Regression model:**





From this Model, Except Northwest and Sex remaining all variables are significant.

We get **75.09%** of variability with Charges and **86%** of Correlation with Charges.

By eliminating insignificant variables it might get increase in variability,doing another model by not considering insignificant variables to know whether it changes or not.

**2nd Regression Model:**





After eliminating previous model insignificant variables, from this model we get **75.05%** Variablity and getting Southwest as a insignificant variable.

**3rd Regression Model:**





By eliminating all the columns of Region, we get **74.9%** of variability with charges and 86.5% of correlation.

It is having small variance of 0.01% from 1st model variance means our preffered model is 3rd which is having only Age,bmi,children,smoker gives the best correlation with Charges.

This is the Best fit model.