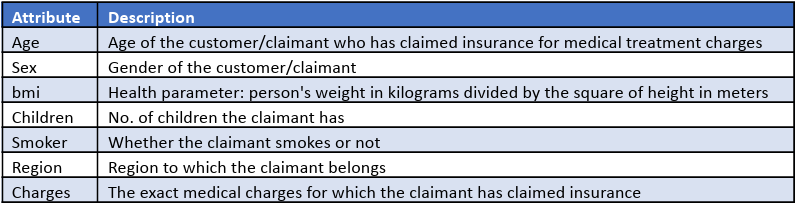
Health Insurance claim

**Problem Statement (Situation):**

“Finding out the health parameters that affect health insurance claims”

An insurance company in the US is reviewing its insurance claims/charges and is trying to do a cause and effect analysis for future business decisions. It has collected data for its customers who have made claims till recent time. The data-points collected are age, gender, bmi, number of children/dependents, smoking habit, region they belong to, charges/bills claimed under the insurance. This analysis would have a bearing on what premium should the company charge a customer availing an insurance policy. The insurance company has collected a dataset of 1338 customers-claims. Please refer to the data dictionary below:

Data Dictionary:



Objective (Task):

● To do a cause and effect analysis on historic-data of insurance claims.

1. Perform the Exploratory Data Analysis on the data.

a) Identify the categorical and continuous variables

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

ii. Charges vs Age

iii. Charges vs BMI

iv. Charges for Smokers vs Non-smokers

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

e) Region-wise charges for smokers vs 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)

2. Edit the data as following, to obtain dummy variables: (5 marks)

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.

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.